Weighting ratings: are people adjusting for bias in extreme reviews?

Neel Ocean

University of Warwick

July 18, 2023

Running head: ADJUSTING FOR EXTREME RATINGS

Author note:

Address: WMG, University of Warwick, Coventry, CV4 7AL, UK.

Email: neel.ocean@warwick.ac.uk

The author thanks Andrew Oswald, Daniel Sgroi, Gordon Brown, Peter Howley, Casey Weaverling, Wändi Bruine de Bruin, Lukasz Walasek, and seminar participants at Warwick and Leeds for their insightful comments. Funding was provided by CAGE and the Behaviour, Brain & Society Global Research Priority at the University of Warwick. Part of this work is based on a chapter from the author's PhD thesis. Ethical approval for data collection was obtained under the University of Warwick DR@W umbrella. Datasets are available at https://www.doi.org/10.17605/osf.io/9zkwn, and code required to run the analyses can be provided on request. Study 1 was not preregistered. Study 2 was preregistered on AsPredicted: https://aspredicted.org/blind.php?x=VP1_FCR. The author has no conflicts of interest to report.

Abstract

The increasing importance of consumer ratings raises the question of whether people adjust for potentially fake or biased extreme opinions when judging products. Two studies tested treatments that trimmed the extremes of rating distributions. Neither removing extreme ratings while preserving the mean, nor flagging suspicious extreme ratings, nor priming individuals about review manipulation significantly affect judged product quality on average. However, judgements for specific distributions may be made less extreme by flagging or trimming. On average, it is difficult to override usage of the mean rating as the strongest proxy for product quality. When a weighted-mean model is fitted, the estimated weighting profile is hump-shaped and asymmetric. Consumers appear to discount 5-star ratings but are particularly susceptible to being misled by disingenuous 1-star ratings. The weights suggest that there is a binary bias with an inflection point at 2-stars for product ratings, meaning that any rating above this broadly sends an equally strong positive signal of quality. Further theoretical work is required to understand how people form weights for ratings, and applied work should continue to search for decision aids that could help consumers to better adjust for review bias.

Keywords: rating distributions, fake reviews, trimming, weighted mean

Public significance statement: This study finds that people interpret product ratings in a binary way. Ratings of 3-stars or above are viewed similarly, whereas 1-star ratings are viewed very negatively. It is difficult to override the mean rating as a proxy for product quality, so mechanisms that flag suspicious reviews may not be effective.

Introduction

People increasingly have to make judgements about the quality of a product or service from numerical user ratings. Reviews and ratings accompany the vast majority of products and services currently available for purchase online, and consequently inform purchasing decisions (e.g. King, Racherla, & Bush, 2014; Ludwig et al., 2013), especially where physical access to products is not possible (e.g. Babić Rosario, Sotgiu, De Valck, & Bijmolt, 2016; Luca, 2016). Their importance has steadily increased due to the fact that an increasing proportion of consumer transactions are made online. For example, the proportion of total retail sales in the U.S. accounted for by e-commerce has more than tripled in a decade, from 4.7% in Q1 2011 to 14.8% in Q1 2021 (U.S. Department of Commerce, 2022).

A reliance on reviews and ratings creates two main problems. First, reviews can be manipulated to intentionally change consumer judgement (Hu, Liu, & Sambamurthy, 2011). Media reports of review manipulation have become widespread, with firms leaving reviews for their own or competing products, and even buying the services of fake reviewers (e.g. Box & Croker, 2018; Brodkin, 2021). Empirical evidence suggestive of falsified negative reviews has been found for competing hotels (Mayzlin, Dover, & Chevalier, 2014), and fake reviews have been shown to have a significant causal effect on Amazon.com product sales (He, Hollenbeck, & Proserpio, 2022). Second, systematic biases in rating behaviour like herding (Sunder, Kim, & Yorkston, 2019) may lead to inaccurate judgements and subsequent poor choices.

In light of these issues, the present study focuses on two broad questions. First, can a simple mechanism be used to "treat" the ratings summary so that people are less likely to follow biased signals in making judgements? Second, how are people judging the quality of a product from numerical ratings summaries, and can we use this information to determine to what degree a disingenuous or biased review may affect an evaluation? To investigate these questions, data from two randomised judgement experiments were collected. The task required participants to independently judge the quality of ten different products or services. The interventions involved either "trimming" the extremes of the ratings distribution in some way, or priming individuals about fake reviews. The trimming of extreme opinions is a common heuristic strategy used to improve performance in the judge-advisor literature, where individuals must determine how best to integrate the information from a number of advisors in order to arrive at the most accurate judgement (e.g. Bonaccio & Dalal, 2006; Yaniv, 1997; Yaniv & Milyavsky, 2007). Subsequently, a model-fitting exercise was carried out using observations

from both experiments in order to understand whether any weighting or adjustment of numerical ratings is taking place, and to what extent (if any) this adjustment process could mitigate for the presence of disingenuous reviews.

A number of studies have investigated how distributional characteristics of ratings such as valence (i.e. mean rating), variance, and volume affect purchase decisions. The general finding from these studies is that higher valence, higher volume, and lower variance have a positive effect on demand, though the variance effect is only true for products with a high rating valence to begin with (Chevalier & Mayzlin, 2006; Chintagunta, Gopinath, & Venkataraman, 2010; Etumnu, Foster, Widmar, Lusk, & Ortega, 2020; Langan, Besharat, & Varki, 2017; M. Sun, 2012). *Extreme ratings* (or *extreme reviews*) are usually defined as ratings at the endpoints of the scale (e.g. 1-star and 5-star ratings). There is both general and context-specific evidence that people focus on extremes. For example, people pay more attention to the endpoints of a distribution in risky choice tasks (Ludvig, Madan, McMillan, Xu, & Spetch, 2018). In a ratings context, Etumnu et al. (2020) estimates that 5-star ratings have the greatest marginal impact on ground coffee sales, relative to ratings of any other score. Extreme ratings are seen as more helpful in some contexts, presumably because they send a less ambiguous signal about quality (Park & Nicolau, 2015), though this may not be universally true for all types of good (e.g. Mudambi & Schuff, 2010). A study on DVDs found that extreme ratings were significantly positively related to review helpfulness, though only when there was high dispersion in ratings (Lee, Lee, & Baek, 2021). In sum, choices are influenced by properties of ratings distributions, and particular attention is likely to be given to extreme ratings.

The aforementioned studies use secondary data on ratings and outcomes. Because of this, they cannot say much about how people judge products from ratings at the individual level. The present study focuses on individual-level judgements by isolating how consumers weigh ratings of different scores from numeric-graphical rating distributions. By doing so, we can estimate how susceptible people would be to disingenuous ratings, particularly at extreme ends of the scale, as well as begin to understand the underlying processes behind the mapping of rating information to product judgements. Recent studies that investigate elements of the judgement process using similar experimental designs are Köcher and Köcher (2021) and Fisher, Newman, and Dhar (2018), who investigated the mode heuristic and the binary bias respectively. The mode heuristic suggests that individuals focus on the rating that is associated with the highest volume of reviews. The binary bias refers to the tendency for people to distill judgements that

can take a range of values into a "good" or "bad" assessment by binning portions of the rating scale. This implies that choices based on ratings distributions are made by comparing the number of good signals to the number of bad signals.

Why should we care about whether people can correct for bias in ratings? False information can influence judgement and lead to suboptimal choices. This is because a false fact or piece of news that is familiar or easy to process is often believed to be true, even in the face of knowledge to the contrary (e.g. Fazio, Brashier, Payne, & Marsh, 2015; Pennycook, Cannon, & Rand, 2018). Fake reviews may themselves have negative psychological implications for consumers aside from poor decision-making; they increase distrust, increase uncertainty, and create discomfort (Wu, Ngai, Wu, & Wu, 2020). Although automated detection methods have been proposed using sentiment analysis and opinion mining (e.g. Lim, Nguyen, Jindal, Liu, & Lauw, 2010; Mukherjee, Liu, & Glance, 2012), it remains difficult to identify and remove all non-genuine reviews (see Liu & Zhang, 2012, for a review). Therefore, fake reviews are likely to remain ubiquitous. Hu, Bose, Koh, and Liu (2012) estimated that 10.3% of products on Amazon.com contain manipulated reviews. Despite this, average ratings influence perceived product quality more than objective signals such as price (De Langhe, Fernbach, & Lichtenstein, 2016), suggesting fake reviews potentially have a large influence on product choice if people do not (or cannot) discount them.

Aside from deliberate manipulation, ratings may be inherently likely to send the consumer a biased signal of true product quality. A subjective evaluation of an objective quantity can be characterised as *true value + systematic bias + random error* (Yaniv & Milyavsky, 2007). Therefore, even if errors in judgement have zero mean, aggregate product ratings may still not represent the true quality of a product if there is a nonzero systematic bias. Ratings could signal true quality if all consumers left a review, or if those consumers that do leave a review are equally likely to "moan" about a bad product as they are to "brag" about a good one (Hu, Pavlou, & Zhang, 2006). Empirically, however, very satisfied or very dissatisfied consumers are more likely to voice their opinions (Koh, Hu, & Clemons, 2010), with around half of the products on Amazon.com having bimodal ratings distributions (Hu et al., 2006). For example, workers with extreme experiences are more likely to share their opinions by leaving voluntary employer ratings (Marinescu, Chamberlain, Smart, & Klein, 2021). Hence, it is plausible to assume that the systematic bias in product ratings is not zero, independent of deliberate manipulation. Furthermore, high ratings are more common than low ratings, creating a J-shaped distribution (Hu,

Zhang, & Pavlou, 2009). Therefore, extreme ratings are more common than moderate ratings and are likely to contain inherent systematic bias.

Can we reduce the impact of extreme opinions? The judge-advisor literature (see Bonaccio & Dalal, 2006) deals with exactly this problem. Its general aim is to determine how individuals (*judges*) can arrive at the most accurate judgement when given advice from multiple parties (*advisors*). Multiple studies have found that trimming outlying opinions and then taking the mean of the remaining opinions is more accurate than relying on the unweighted or untreated mean rating (Lyon, Wintle, & Burgman, 2015; Yaniv, 1997; Yaniv & Milyavsky, 2007). Trimming is readily used in sporting events where performance is based on subjective scores from judges who are inherently biased (e.g. Heiniger & Mercier, 2018). To maximise accuracy, trimming should be consensus-based rather than ego-based (Yaniv & Milyavsky, 2007). That is, judges should not trim outliers relative to their own starting judgements, but should trim outliers relative to the average of the judgements of all advisors. Judges tend to use simple heuristics (like taking the median) when they are less experienced, but may start to put more weight on outlying opinions the more experienced they become because they may be more likely to observe examples of outliers reflecting the truth (Harries, Yaniv, & Harvey, 2004). These insights suggest that an externally imposed trimming method could lead to more accurate judgements by consumers, who would otherwise be disproportionately swayed by the relatively high volume of extreme ratings.

The main contributions of the present study are that: (i) it tests the impact of an externally imposed trimming mechanism on ratings that, to the best of the author's knowledge, has not been utilised in the context of online reviews and ratings; and (ii) it estimates implicit weights placed on ratings of different scores so that we can predict the marginal effect of an additional rating on quality judgements. The findings suggest that an externally imposed trimming mechanism that changes the ratings distribution is ineffective on average unless it also changes the mean rating substantially. Highlighting the trimmed ratings without changing the mean is also ineffective in changing judgements on average, although it may be effective in situations where the distribution of ratings would not visibly change after trimming. Finally, priming people with information about review manipulation does not change judgement at all. Regression estimates of weights on ratings suggest that both a negativity bias and a form of binary bias are present when people translate ratings to judgements.

Hypotheses

We first discuss why trimming extreme product ratings could help to reduce bias in judgements. Upon seeing a rating distribution and the mean rating, the most naive proxy for product quality would simply be to take the mean as the sole indicator. Taking the mean is also sometimes considered a normative way to integrate opinions for more accurate conclusions (Yaniv, 1997). It follows that for a disingenuous review to have the largest impact on a product's evaluated quality, it must change the mean by as much as possible in the intended direction. Therefore, if an individual or firm wanted to positively (negatively) influence how a product is perceived, then their best response would be to give the highest (lowest) possible rating to it.

Our context has parallels to the judge-advisor literature. Work stemming from Yaniv (1997) has focused on weighting and trimming heuristics that people use to integrate advice from multiple sources to arrive at the most accurate judgement. Using a *trimmed* mean rather than the untreated mean minimises judgement error as long as the distribution of the error is symmetric, centred around zero, and has relatively thick tails (Yaniv, 1997). Since any subjective judgement of a latent objective value contains a systematic bias component (Yaniv & Milyavsky, 2007), the untreated mean alone is not an effective judgement heuristic because the central limit theorem would not apply. This is because taking a large sample of opinions will eliminate noise but not bias. Trimming reduces systematic bias from outlying opinions, bringing the estimated judgement closer to the true value.

In our context, the systematic bias component comes from the unconscious bias of reviewers, as well as deliberately disingenuous extreme ratings that may be completely independent of true product quality. A more sophisticated consumer might anticipate that extreme ratings are akin to extreme opinions, and discount these extreme opinions using their own trimming strategy (Harries et al., 2004). However, a robust finding from the judge-advisor literature is that individuals are egocentric in their application of any discounting (Bonaccio & Dalal, 2006). This means that disproportionate weight is given to their own prior opinion even if they do appear to discount some extreme opinions (Yaniv & Milyavsky, 2007). Instead, a consensus-based trimmed mean is required for accurate judgements. This means the outliers that are trimmed from the distribution are objective outliers in terms of the distribution of the opinions of advisors, and not outliers based on the view of the decision-maker alone (Yaniv & Milyavsky, 2007). Translating this to a typical five-point online rating scale, we could externally impose trimming or censoring of 1-star and 5-star ratings to reduce the likelihood of being misled by

outlying ratings, either from the tendency to moan/brag or from fake reviews. This should minimise the adverse impact of the consumer applying their own flawed egocentric heuristics to the ratings distribution.

The above arguments provide justification for imposing trimming on a ratings distribution. How would individuals judge the quality of a product from its ratings distribution once it has been trimmed? In an online product evaluation setting, we cannot determine the inherent quality of a product. Therefore, we cannot measure normative accuracy gains from applying an aggressive trimming to the ratings distribution. Usually, trimming would alter the mean rating and so individuals would be likely to base their evaluations on this new trimmed mean. However, we could determine whether an externally imposed trimming treatment changes the judgement process by keeping the mean rating unchanged after trimming. If individuals used only the mean rating to judge quality then there should be no change in judgement before and after a trimming that preserves the mean. On the other hand, if individuals did pay attention to the shape of the distribution, then they should notice a reduction in variance at the extremes. In this case, trimming should steer them away from more extreme judgements about product quality. A lower variance in ratings can have a positive effect on demand (Langan et al., 2017), and this suggests that distributional characteristics are likely to be taken into account by decision-makers. Therefore the first hypothesis is as follows:

Hypothesis 1 *Individuals evaluate the quality of goods differently on average when 1-star and 5-star ratings have been trimmed (controlling for the mean rating).*

It is likely that judged quality would increase on average as we remove extreme ratings due to the finding by Langan et al. (2017), though the direction is somewhat ambiguous because higher variance can increase sales when other characteristics are accounted for (Etumnu et al., 2020).

The second stage of the investigation is based on the assumption that consumers may be applying weights to ratings in order to mitigate for the effects of review bias. Here, we are interested in determining what exactly these weights are. We are particularly interested in the weights at the extremes because we could think of weighting as a self-imposed form of trimming. Existing cognitive accounts such as range-frequency theory (Parducci, 1968) and decision-by-sampling (Stewart, Chater, & Brown, 2006) explain judgements based on the rank or range of a signal in the context of the distribution of signals in memory and/or the choice environment (Brown & Matthews, 2011). These models are difficult to directly apply to the current

context, because the range of the ratings scale is always constrained. Therefore, for example, a very high average rating implies a relatively limited range of distributional shapes. Additionally, these studies usually present stimuli sequentially and rely on participants to infer the distribution from memory, whereas individuals generally see the entire ratings distribution.

The mean rating has been found to be the strongest single signal of sales and quality, though it is true that people are also influenced by characteristics of the ratings distribution (e.g. De Langhe et al., 2016; Etumnu et al., 2020). Evidence on ensemble statistics suggests that people can recall summary statistics like the mean, maximum, and minimum when shown a set of stimuli (e.g. Ariely, 2001; Brady & Alvarez, 2011; Putnam-Farr & Morewedge, 2021). While the present context is slightly different because ratings distributions are essentially summary statistics of individual reviews, individuals may be basing their initial judgements on the mean, and then adjusting this judgement based on key distributional statistics particularly those at the extremes (i.e. corresponding to a maximum and minimum). How might they be making this adjustment? If people are mitigating for extreme opinions then 1-star ratings should be given a weight greater than 1 (assuming they are overly pessimistic indicators of true quality) and 5-star ratings should be given a weight less than 1 (assuming they are overly optimistic indicators of true quality).

To formalise this, suppose 1-star, 2-star, 3-star, 4-star, and 5-star reviews are given implicit weights w_1, w_2, w_3, w_4, and w_5 respectively. Then the null hypothesis to test each weight against is $w_r = 1$, where $r \in \{1, 2, 3, 4, 5\}$ is the corresponding rating. This is because if all weights were equal to 1, the weighted mean would simply equal the mean. The weighted mean is given by:

$$\mu_w = \frac{1}{N} \sum_{r=1}^{5} w_r(rn_r) \tag{1}$$

where n_r is the number of reviews with a rating of r, $N = \sum_{r=1}^{5} n_r$ is the total number of reviews, and w_r is the weight attached to a review of score r. We would expect that w_1 and w_5 are further from 1 than w_2, w_3, and w_4 because we are assuming that it is extreme opinions that people are most concerned with adjusting to improve the accuracy of their judgements.

Given our focus on extreme reviews, what would we expect the weights on 1-star and 5-star ratings to be? As a benchmark, assume the consumer updates in an "approximately Bayesian" way. The consumer begins with the prior belief that ratings should be taken as-is, with the mean serving as an unbiased estimate of expected quality. However, a consumer with experience of

online ratings will see the J-shaped distribution (Hu et al., 2009) and thus the tendency for reviewers to represent extreme viewpoints. Knowing that most products are unlikely to be of either minimum or maximum quality, the posterior would reduce the likelihood that a 5-star rating is signalling a maximum quality product, and also reduce the likelihood that a 1-star rating is signalling a minimum quality product. If 5-star ratings are believed to be an inflated proxy for true quality, then $w_5 < 1$. The lowest plausible value for w_5 is 0.6, since 5-star ratings would then be perceived equivalently to 3-star ratings. Below this, a 5-star rating would act as a signal of poor quality, which would be implausible. Therefore, if consumers were to apply a partial adjustment of 5-star ratings towards the midpoint of the scale in response to perceived bias, we would expect that $w_5 \in (0.6, 1)$. At the other end of the scale, if consumers were under the impression that 1-star ratings were an underestimate of true quality, they should place a weight on them that is greater than 1. If $w_1 = 3$, then 1-star ratings would be perceived equivalently to 3-star ratings, which would remove all negative valence from them. Therefore, $w_1 \in (1, 3)$ represents the range of weights that the consumer would realistically apply to reduce the negative valence of 1-star ratings.

There is little reason to believe that 5-star ratings would be given a weight greater than 1. However, literature on negativity bias (e.g. Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001) suggests that the weight placed on 1-star ratings could act in the opposite direction. Baumeister et al. (2001) explain that the impact of bad events (or stimuli) are longer lasting and stronger in magnitude than the impact of good events. Previous work has found a positive-negative asymmetry when forming impressions or judgements of others (e.g. Peeters & Czapinski, 1990), such as negative ratings reducing trust in an eBay seller more than a positive rating increases it (Ba & Pavlou, 2002). In line with this research, we may actually expect in practice that the weight placed on 1-star ratings is less than 1, and furthermore, is stronger (i.e. further from 1) than the weight placed on 5-star ratings. In summary, the hypotheses for weights are:

Hypothesis 2a *Individuals adjust for bias in 5-star ratings so that* $w_5 \in (0.6, 1)$

AND

Hypothesis 2b *Individuals adjust for bias in 1-star ratings so that* $w_1 \in (1, 3)$

OR

Hypothesis 2b Alt *Individuals do not adjust for bias in 1-star ratings (due to negativity bias) so that* $w_1 \in (0, 1)$ *and* $w_1 < w_5$

Finally, we form ancillary hypotheses that aim to test how the type of product moderates each of the previous hypotheses. Nelson (1970) classified products into *search goods* and *experience goods*. Though any good can have both search and experience characteristics, most goods are likely to have a dominant type (Klein, 1998; Nelson, 1974). Search goods have well-defined objective attributes, and therefore their value can be compared relatively easily by comparing these attributes (e.g. a camera). Purchase of a search good is, in general, not necessary to evaluate it relative to its substitutes. In contrast, experience goods lack such comparable attributes. Their value depends largely on subjective experience (e.g. a restaurant), for which purchase is usually required to obtain.

The search vs experience distinction is closely related to other product classifications used in the literature. Products can have vertical attributes that are quality-based or horizontal attributes that are taste-based (Spiller & Belogolova, 2017). There is a close correspondence between vertical attributes and search characteristics, and between horizontal attributes and experience characteristics. However, there are some subtle differences. For example, a hotel is a classic example of an experience good because one has to stay there to evaluate it accurately. While hotels are predominantly characterised by taste, they are also vertically-differentiated (e.g. by the services they provide). Yet another related but subtly different product classification distinguishes experiential purchases from material purchases. Gilovich and Gallo (2020) explain that the main difference between experiential purchases and experience goods is that the search vs experience distinction is focused on how individuals determine the quality of that good, rather than making an assertion about its inherent properties. For example, a book is an experience good because its quality cannot be evaluated by objective attributes, but it is not an experiential purchase because it is still a material product.

We would expect products with strong horizontal / experiential / experience characteristics (i.e. predominately taste-based goods) to have a naturally large spread of ratings. Ratings for taste-based experience goods by definition will contain a large element of personal preference, and so their distribution is more likely to be reflective of tastes in the sample than providing a distribution around a true judgement of objective quality. Furthermore, some experience goods may not even be considered to have an objectively defined true quality. Therefore, it is likely that ratings for experience goods and ratings for search goods are interpreted differently by con-

sumers. Indeed, consumers spend more time and effort researching reviews and information for experience goods than search goods (P. Huang, Lurie, & Mitra, 2009). Reviews for experience goods require fewer "helpfulness votes" to be considered useful by the consumer than reviews for search goods (X. Sun, Han, & Feng, 2019) and they communicate more in terms of unique sentiment (Guha Majumder, Dutta Gupta, & Paul, 2022). Thus, one would expect extreme opinions for experience goods to be seen as *more valid* than extreme opinions for search goods. However, extreme ratings for experience goods are considered as *less helpful* than moderate reviews depending on the product in question (Mudambi & Schuff, 2010).

Therefore, because of the difference in the interpretation and helpfulness of extreme ratings for search goods relative to extreme ratings for experience goods, we would product type to moderate the effect of trimming treatments. By extension, we would also expect that the weight placed on extreme ratings would differ between search and experience goods. How they would differ is less clear, however. Extreme ratings for search goods may be seen as a sign of systematic bias and so be discounted more heavily. With this logic, we would expect weights on 1-star and 5-star to be closer to one for experience goods. However, the perceived low helpfulness of extreme ratings for experience goods (but not search goods) may mean that extreme ratings are actually discounted more heavily for experience goods. In this case, weights on 1-star and 5-star ratings should be closer to one for search goods. In summary:

Hypothesis 3a *Holding mean rating constant, trimming 1-star and 5-star ratings has a different impact on judged product quality depending on whether the product is an experience good or a search good.*

Hypothesis 3b *The weights placed on 1-star and 5-star ratings will be significantly different for experience goods, relative to search goods.*

Method

Transparency and openness

Study 1 was not preregistered and data were collected in June 2016. Study 2 was pre-registered on AsPredicted (https://aspredicted.org/blind.php?x=VP1_FCR), and data were collected in July 2022. Datasets for both studies are available on the OSF at: https://www.doi.org/10.17605/osf.io/9zkwn, as well as code to reproduce the analyses and full Qualtrics survey scripts

(Ocean, 2023). Data were analysed using Stata (StataCorp, 2013), and all plots were generated using the ggplot2 package in R (R Core Team, 2023). Sample sizes are justified in the Participants sections for each study, and the study methodology is described in detail below. Ethical approval for data collection was obtained from the University of Warwick's Humanities and Social Sciences Research Ethics Committee under the DR@W umbrella. The article adheres to the recommendations in the APA JARS-Quant.

General design of studies

In order to test the hypotheses, two separate online studies were run. Both studies used a similar design, but tested slightly different treatments. The general design is outlined below, followed by details specific to each study.

Individuals were asked to rate the quality of a series of products and services that they were shown, based only on the quantitative summary information of ratings that is commonly found on online product pages. Similarly to Hu et al. (2011), the present study uses perceived quality as an outcome, rather than purchase intentions. From a judgement perspective, understanding product quality arguably has wider implications than purchase intention alone. Theoretically, it is useful to understand how informative reviews are as a signal of the "true" value of an economic good (e.g. see W. Dai, Jin, Lee, & Luca, 2018; Song, Park, & Ryu, 2017), because this can help to inform the design of review platforms that seek to generate unbiased signals of product value. More practically, a product that is considered high quality may still not be purchased (e.g. due to tastes), but might instead be recommended to other potential consumers, or may influence the purchase of a different good or service from the same producer. Participants were also asked for their maximum willingness to pay (WTP) for each item, primarily to provide better separation between product quality in an objective sense and the subjective value to a particular individual.

Ten items and their ratings information were selected from the websites Amazon.co.uk and TripAdvisor.co.uk. Each item on these websites is rated on a 1-5 scale by registered users who wish to leave a review (these are referred to as *star ratings* in this paper for convenience although TripAdvisor uses circles in place of stars). Because anyone with an account on these websites can review any product, and there is no restriction to who can create an account, fake reviews are likely to exist. For each item, a brief title was displayed, along with one or two images. Branding was stripped from the images in order to minimise bias caused by brand loyalty. For each good,

the distribution of ratings was shown along with the mean score in a similar format to what is shown to website visitors. Amazon shows the mean both numerically and visually, whereas TripAdvisor only shows the mean visually. This presentational difference was preserved in the experiment to maximise external validity. Review information was presented in a format which was in keeping with the style of the source website (see Figure 1 for an example). Participants were asked to report the quality of each item on a 0-100 scale. This scale was chosen for two reasons: to allow for a fine-grained judgement, and to make it less likely that participants would reflexively copy the mean rating.

Of the ten items selected, five were search goods and five were experience goods. This allows us to test Hypotheses 3a and 3b. For each type of good, products were selected in order to cover the following criteria: highly-rated (i.e. positive valence) with a high overall number of reviews; poorly-rated (i.e. negative valence) with a high overall number of reviews; highly-rated with a low overall number of reviews; poorly-rated with a low overall number of reviews; and middle-rated (neutral valence) with a roughly even split of bottom and top reviews (i.e. a U-shaped score distribution). Highly (poorly) rated in this context refers to mean scores that are above (below) the mid-point of the review scale. A low volume of reviews was defined as $N < 55$ for search goods and $N < 32$ for experience goods, whereas a high volume of reviews means $N > 100$. The lower thresholds were based on the 95th percentiles of the total review volume for electronics and books on Amazon from an existing dataset (McAuley, Pandey, & Leskovec, 2015; McAuley, Targett, Shi, & van den Hengel, 2015). This spread of ratings distributions was chosen in order to represent those most commonly found online. See Table 1 for a full list of product types and characteristics.

Following the experimental task, participants were asked to complete the 20 item mini-IPIP personality inventory (Donnellan, Oswald, Baird, & Lucas, 2006; Goldberg et al., 2006), based upon the Big Five factors. These personality measures were collected to control for the effects of individual differences on product evaluation. A full list of personality items can be found in Table S1 in the online Supplemental Material. The order of items was randomised per-subject and responses were given on a seven-point Likert scale. Information on a number of demographic variables was also collected (age, sex, marital status, employment status, education, income) after the main task had been completed.

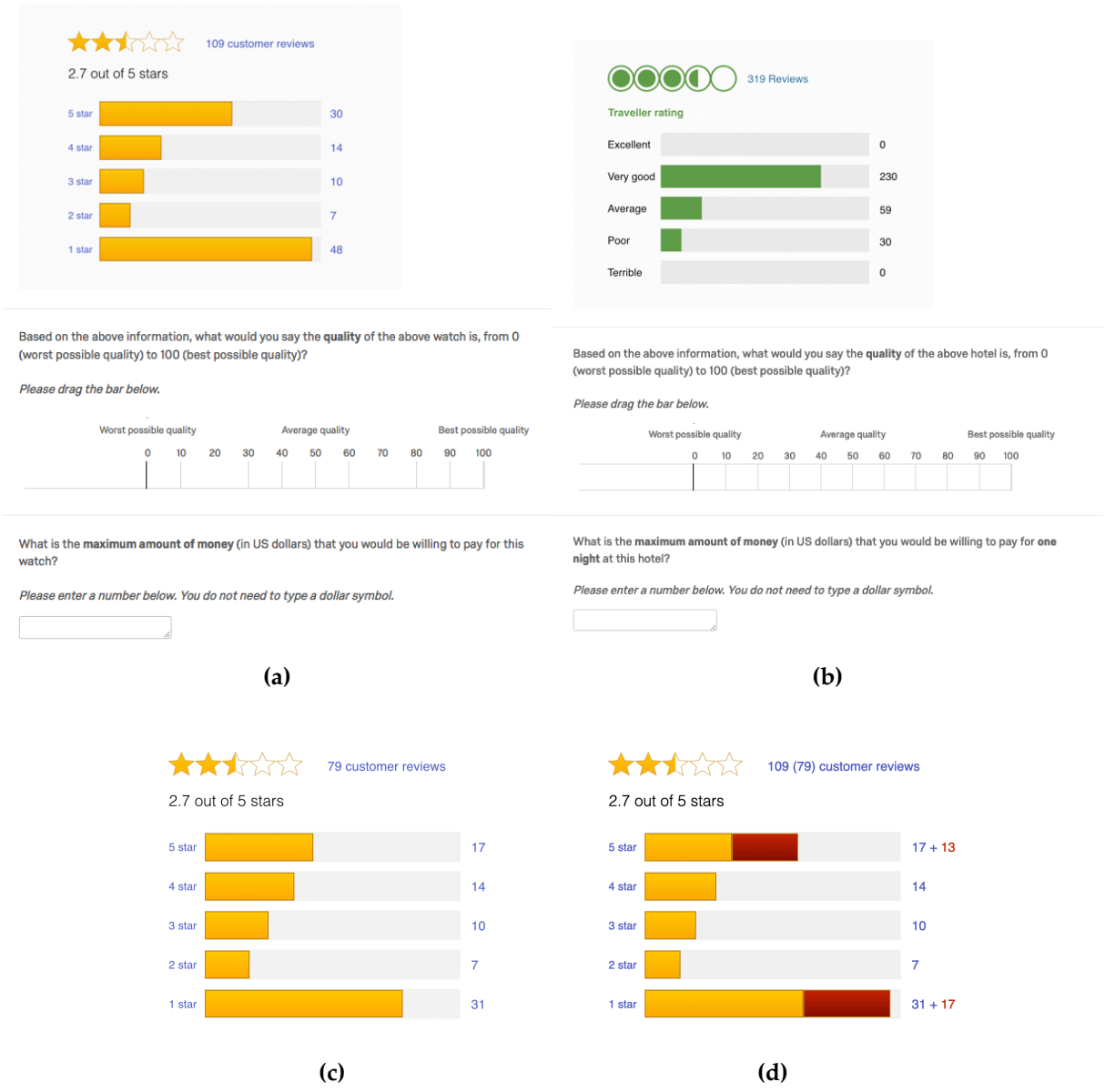


Figure 1: Example of the information and questions shown to a participant for a good. (a) The control condition (i.e. unadjusted score distribution from Amazon.co.uk) for good 1 - a smartwatch. (b) The no-extreme (T2) treatment for good 7 - a hotel, from TripAdvisor.co.uk. Note that product images were displayed above the rating distributions but are omitted from this figure. (c) The mean-preserving treatment (T1) for the product distribution in panel (a), as used in Study 1. (d) The treatment used in Study 2, where instead of removing ratings entirely, they were highlighted in red.

Table 1: Summary of the 10 products used for both studies

Good	Description	Type of good	Treatment	Mean	# 1*	# 2*	# 3*	# 4*	# 5*	Total
1	Smartwatch	Search	Control	2.7	48	7	10	14	30	109
	(poorly rated, high N)		T1	2.7	31	7	10	14	17	79
	Amazon		T2	3.2	0	7	10	14	0	31
2	Smartphone	Search	Control	4.4	23	32	38	94	370	557
	(highly rated, high N)		T1	4.4	5	32	38	94	276	445
	Amazon		T2	3.4	0	32	38	94	0	164
3	Headphones	Search	Control	2.5	9	6	2	2	5	24
	(poorly rated, low N)		T1	2.5	4	6	2	2	2	16
	Amazon		T2	2.6	0	6	2	2	0	10
4	LCD TV	Search	Control	4.5	2	2	3	6	39	52
	(highly rated, low N)		T1	4.5	1	2	3	6	32	44
	Amazon		T2	3.4	0	2	3	6	0	11
5	Laptop	Search	Control	3.2	86	23	29	35	102	275
	(split opinion)		T1	3.2	17	23	29	35	21	125
	Amazon		T2	3.1	0	23	29	35	0	87
6	3-star Hotel	Experience	Control	2.2	485	184	248	129	64	1110
	(poorly rated, high N)		T1	2.2	346	184	248	129	5	912
	TripAdvisor		T2	2.9	0	184	248	129	0	561
7	4-star Hotel	Experience	Control	4.1	18	30	59	230	206	543
	(highly rated, high N)		T1	4.1	5	30	59	230	164	488
	TripAdvisor		T2	3.6	0	30	59	230	0	319
8	Programming book	Experience	Control	2.5	14	2	4	5	5	30
	(poorly rated, low N)		T1	2.5	9	2	4	5	2	22
	Amazon		T2	3.3	0	2	4	5	0	11
9	Parenting book	Experience	Control	4.3	3	3	3	5	34	48
	(highly rated, low N)		T1	4.3	1	3	3	5	24	36
	Amazon		T2	3.2	0	3	3	5	0	11
10	Restaurant	Experience	Control	2.9	48	24	21	23	43	159
	(split opinion)		T1	2.9	18	24	21	23	15	101
	TripAdvisor		T2	3.0	0	24	21	23	0	68

Notes: T1 = *Mean-preserving* treatment, T2 = *No-extreme* treatment. # 1* = the number of 1-star ratings, # 2* = the number of 2-star ratings, etc. T1 has reduced numbers of 1-star and 5-star ratings, whilst keeping the mean identical to the original distribution. In T2, all 1-star and 5-star ratings are removed from the original distribution. Control, T1, and T2 distributions were used in Study 1. Study 2 showed either the control distribution, or a treatment where the full distributions were shown but the difference between the control and T1 distributions were highlighted in red to represent suspicious reviews.

Study 1

Participants A total of 501 participants were recruited in summer 2016 via Amazon MTurk and paid $1.50 to fill out an online study via Qualtrics, so that there would be approximately 167 observations per product-condition combination across the three different distributional conditions. A power calculation assuming a moderate effect size (Cohen's $d = 0.5$) suggested a minimum required sample size of 105 per group for a two-tailed t-test for the difference between two independent means with $\alpha = 0.05$ and power = 0.95. The minimum number of observations that were actually obtained for a single product-condition combination was 165 and the maximum was 169. No responses were excluded. Mean completion time was 12.76 minutes. Mean age was 37.7, 51.7% were male, and 69.5% had at least an undergraduate degree.

Procedure In order to test Hypothesis 1, we require treatments that manipulate the number of 1-star and 5-star ratings shown to respondents. Each individual was asked to rate all 10 items (in a random order), but with equal probability saw one of three possible conditions for each item. The *control* condition showed all the information as described above with the original ratings distribution. The first trimming treatment (T1), referred to as *mean-preserving*, removed *some* 1-star and 5-star ratings from the distribution, whilst keeping the mean unchanged. This was done by writing a small algorithm that computed all mean-preserving treatments to a specified number of decimal places. From this list, an appropriate distribution was selected – one that (as far as possible) was not too similar to the original distribution. If individuals were only basing their quality judgements on the mean, there should be no difference in reported quality between T1 and the control group. T1 reduces the variance of a ratings distribution by narrowing only at the extremes, steering judgements away from the extremes. An example of this can been seen by comparing panels (a) and (c) in Figure 1. The second trimming treatment (T2), referred to as *no-extreme*, removed *all* 1-star and 5-star ratings. This represents a more traditional example of the trimmed mean, applied in a heavy-handed way so that all extreme opinions are removed. Doing this necessarily changes the mean, but we can use regression analysis to test whether there is a treatment effect from reducing extremes after controlling for the mean rating. These treatments abstractly represent a scenario in which a trimming mechanism is applied automatically by a website as a decision aid to mitigate for review bias. The rating distributions for each product and treatment are shown in Table 1.

Study 2

Participants The aim was to collect 200 observations per group across four groups (i.e. 800 in total) so that there was a comfortable level of power to be able to test for treatment effects (based on the same calculation as for Study 1), and enough of a buffer to account for a loss in observations due to failing the manipulation check. In total, 803 participants were recruited via Prolific Academic in summer 2022 and paid £2.25, with no special restrictions other than requiring that participants were fluent in English. Both studies employed an attention check on the instructions page which required participants to click on a specific area of the screen before they were allowed to proceed. A total of 28 people failed the manipulation check and were excluded from all further analyses, leaving 775 usable responses. Of these, 198 participants were in the *no-priming, no-flagging* condition (i.e. the control), 191 were in the *priming, no-flagging* condition, 198 were in the *no-priming, flagging* condition, and 188 were in the *priming, flagging* condition. Mean completion time was 9.4 minutes. Mean age was 29.6, 46.1% were male, and 74.2% had at least an undergraduate degree.

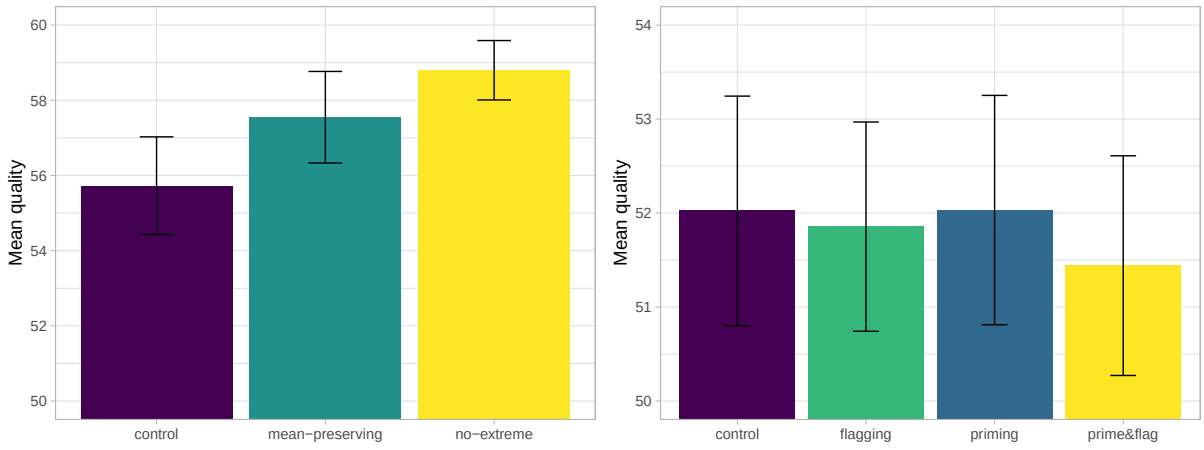
Procedure Whereas Study 1 aims to test the sensitivity of individuals to extreme ratings without mentioning fake reviews, Study 2 tests sensitivity to extreme ratings when the presence of disingenuous reviews was made explicit. The study employed a 2x2 design in order to test two separate treatments: an extreme review *flagging* treatment, and a fake review *priming* treatment. Unlike Study 1, participants were assigned to a particular condition for the entire experiment in Study 2 rather than seeing different treatments on a per-product basis. The flagging treatment trims ratings in exactly the same way as the mean-preserving treatment (T1) from Study 1. However, rather than removing 1-star and 5-star ratings and changing the distribution, the flagging treatment in Study 2 colours a portion of the bars in red. The red portion represents the ratings removed from the control distribution to get to the T1 distribution from Study 1 (see Figure 1d). Therefore, participants would see the control distribution from Table 1, but the reliable (yellow) portion of this distribution would be identical to the distribution in T1. In the priming treatment, participants were shown a screen before the rating task that draws attention to the possibility of fake reviews, along with screenshots and a small excerpt of two news articles about fake reviews (https://arstechnica.com/tech-policy/2021/02/fake-amazon-reviews-still-sold-in-bulk-it-costs-10900-for-1000-reviews/ and https://www.bbc.co.uk/news/technology-52771913). A simple manipulation check question was

shown immediately after the priming screen to ensure participants did not skip or ignore it.

Results

Main treatment effects for both studies

Figure 2 plots mean quality judgements across treatments for both studies. In Study 1, we see that both the mean-preserving ($p = .045$, Cohen's $d = .0694$) and no-extreme ($p = .0001$, Cohen's $d = .137$) treatments increase average perceived quality significantly. In Study 2, neither the flagging ($p = .842$, Cohen's $d = .00633$) nor the priming ($p = .993$, Cohen's $d = .000298$) treatments had significant effects. Although combining them was more effective in changing quality judgements, the difference was not significant relative to the control ($p = .500$, Cohen's $d = .0217$). However, these average treatment effects do not control for individual or product characteristics. Therefore, we next analyse the robustness of the treatment effects by fitting regression models.



(a) Study 1 **(b)** Study 2

Figure 2: Mean reported quality pooled for all ten products, with 95% confidence intervals. Study 1 had at least n=165 responses for each product-treatment combination. Study 2 had at least n=188 in each treatment group.

Regression analyses for average treatment effects

We estimate the following model to test the significance of average treatment effects:

$$Quality_{i,g} = \beta_0 + \beta_1 T_{i,g} + \beta_2 \mu_{i,g} + \beta_3 X_{i,g} + \beta_4 N_{i,g} + \epsilon_{i,g} \tag{2}$$

where: i is the individual, $g \in \{1, 2, ..., 10\}$ is the good, T is a vector of treatment dummies, μ is the mean rating (to 1 d.p.), X is a set of variables including personality, demographics, and additional dummies for each good or type of good, and N is the volume of reviews. The X variables are not necessary for determining treatment effectiveness, but were added to basic regressions as a check for robustness. One observation in the data corresponds to the combination of a product and an individual, so that the total number of observations in each regression is ten times the number of individuals included.

A number of methods can be used to estimate equation (2). Multilevel (or mixed) modelling (e.g. Gelman & Hill, 2006), whilst often favoured in some literature, is not necessary in this case as we are not interested in the upper level of the hierarchy, i.e. specifically measuring the differences between individual goods. Additionally, the goods selected are not a random sample of all possible goods, which violates one of the assumptions required for multilevel modelling. Instead, differences between our grouping variable (good) represents "noise" that we must take into account (F. L. Huang, 2016). Using OLS with regular standard errors would result in incorrect inferences, since we would expect error terms to be correlated within each good. This would lead to misleadingly small standard errors and p-values (Cameron & Miller, 2015). However, one can correct for this by using clustered robust standard errors. A regular between-subject experiment where each individual is randomly assigned to one condition (such as Study 2) technically does not require clustering. However, because Study 1 is randomised at the product level (i.e. the distribution for each item an individual sees is drawn randomly from the three conditions, independently of the other items), we should cluster by good (Abadie, Athey, Imbens, & Wooldridge, 2023).

As a robustness check, equation (2) was estimated using Study 1 data using the following alternative approaches: fixed effects and random effects panel models with robust standard errors; a multilevel mixed model with random intercept and slope parameters for the treatment dummy; and OLS with standard errors clustered by good (see Table S2 in the Supplemental Material). There were no meaningful differences in estimates between these approaches. A slight difference was observed in estimates of the treatment effect when using fixed or random effects and grouping by good rather than by person. However, we can obtain virtually identical estimates by including dummy variables for each good in an OLS regression. Therefore, OLS estimates with standard errors clustered by good are presented as the preferred method of estimation in subsequent analyses for simplicity.

Study 1 regression analyses The results from regression analysis on Study 1 data are shown in Table 2. In specification (1), only the treatment dummies, the mean rating, and total number of reviews are included. Both T1 ($\hat{\beta} = 2.03$, s.e. $= 1.201$) and T2 ($\hat{\beta} = 7.93$, s.e. $= 1.760$) resulted in higher judged quality. However, this effect was only statistically significant for T2 unless we cluster by person using a fixed-effects or random-effects model, where it was also significant for T1 (see Supplemental Material Table S2, regressions 4 and 6).

Specification (2) in Table 2 adds dummies to control for product-specific effects. This reduces the treatment effects for both T1 ($\hat{\beta} = 1.02$) and T2 ($\hat{\beta} = 5.24$), and T2 no longer has a statistically significant effect on quality evaluation. However, estimating a similar model using random-effects estimation finds a significant treatment effect for T2 with a similar effect size (see Supplemental Material Table S2). Specification (3) in Table 2 tests the robustness of specification (2) by adding Big Five personality measures and a set of demographic variables (age, sex, income, employment status, relationship status, and education). In theory, there should be little difference between the estimates in specifications (2) and (3), since individuals were randomly assigned to treatments. Indeed, R^2 increases only by 0.02; adjusted R^2 increases only by 0.01; and though the AIC criterion marginally prefers specification (3), the BIC criterion prefers (2) to (3). Therefore, personality and demographics do not mediate treatment effects. When WTP was used as the dependent variable instead of quality, there were no significant differences across treatments (see Supplemental Material Table S3). Individual differences explain more of the variation in WTP than either the mean rating or treatment dummies. The likely reason for this is that WTP captures a large element of personal preference for each specific good (see Supplemental Material Figure S1).[1]

In summary, a t-test suggests a significant effect on judged product quality when we trim 1-star and 5-star ratings and keep the mean rating constant. However, the significance disappears after we control for product-specific effects in a regression model. This means that we cannot be certain that Hypothesis 1 holds, though trimming clearly does broadly affect judgements. The models show that people put a large amount of weight on the mean rating in judging product quality, and that any impact from reducing the variance at the extremes is small.

To test Hypothesis 3a, we estimate specification (4) in Table 2. This is the same as specification (1), but includes a full set of interaction terms between treatments and a dummy variable indicating whether a good was an experience good (goods 6-10) or a search good (goods 1-5).

[1]This was supported by a number of free-text explanations collected at the end of the task.

The interaction terms are positive but not significantly different from zero. Adding product dummies and individual differences in regressions (5) and (6) does not substantially change estimates of the interaction terms relative to regression (4). Hence, the results from Study 1 suggest that Hypothesis 3a does not hold. That is, the impact of the trimming treatment is not moderated by whether a product is an experience good or a search good.

Table 2: How treatments affect average quality ratings in Study 1

	Dependent variable: Judged product quality					
	(1)	(2)	(3)	(4)	(5)	(6)
Mean-preserving treatment (T1)	2.032	1.022	1.018	1.514	0.567	0.566
	(1.201)	(1.280)	(1.274)	(1.873)	(1.610)	(1.586)
No-extreme treatment (T2)	7.929***	5.244	5.244	7.208**	4.823	4.897
	(1.760)	(3.285)	(3.288)	(2.746)	(3.975)	(3.983)
Mean rating	26.71***	26.13***	26.07***	26.73***	26.06***	26.03***
	(1.179)	(2.374)	(2.365)	(1.051)	(2.493)	(2.483)
Total number of reviews (N)	0.00364	-0.0122	-0.0121	0.00314	-0.012	-0.0119
	(0.00232)	(0.00869)	(0.00860)	(0.00243)	(0.00845)	(0.00836)
Experience good				0.211	4.412**	4.470**
				(2.066)	(1.392)	(1.393)
T1 × Experience good				0.98	0.932	0.92
				(2.489)	(2.252)	(2.219)
T2 × Experience good				1.291	0.88	0.721
				(3.308)	(3.501)	(3.515)
Dummies for each good	No	Yes	Yes	No	Yes†	Yes†
Demographic and Big Five controls	No	No	Yes	No	No	Yes
Constant	-34.25***	-30.40***	-33.01***	-34.30***	-29.93***	-32.63***
	(4.716)	(8.392)	(9.150)	(4.158)	(8.745)	(9.500)
Observations	5010	5010	5010	5010	5010	5010
R^2	0.660	0.678	0.680	0.661	0.678	0.680
Adjusted R^2	0.660	0.678	0.679	0.660	0.678	0.679
AIC	40444	40164	40144	40442	40166	40143
BIC	40477	40190	40202	40495	40205	40201

Notes: All regression specifications are estimated using OLS with standard errors clustered by good (i.e. 10 clusters in total). Standard errors are shown in parentheses. The set of demographic variables includes: age, sex, income, employment status, relationship status, and highest level of education. The AIC and BIC are tests of model fit, where the minimum value suggests the preferred model. † The dummy for good 10 was omitted from (5) and (6) due to multicollinearity. *** $p < .01$, ** $p < .05$.

Study 2 regression analyses Table 3 repeats the regressions shown in Table 2, but for Study 2. The results suggest that neither the priming treatment nor the flagging treatment change quality judgements on aggregate. The estimated average treatment effect of the flagging treatment is also lower in magnitude than for T1 in Study 1 in regressions (1)-(4). This suggests that people are not completely dismissing the flagged (red) portions of the rating distribution. The interaction term between priming and flagging was also not significantly different from zero, suggesting that the addition of a priming screen about fake reviews does not significantly in-

crease the effectiveness of the flagging treatment. Furthermore, regressions (5) and (6) show that whether or not a product is an experience good does not moderate treatment effectiveness. This supports the results from Study 1, which suggest that Hypothesis 3a does not hold.

Table 3: How treatments affect average quality ratings in Study 2

	Dependent variable: Judged product quality					
	(1)	(2)	(3)	(4)	(5)	(6)
Priming	0.00818	0.00818	-0.177	-0.129	-0.129	-0.315
	(0.604)	(0.604)	(0.589)	(0.538)	(0.538)	(0.526)
Flagging	-0.168	-0.168	-0.394	-0.765	-0.765	-0.993
	(1.536)	(1.537)	(1.580)	(1.984)	(1.984)	(2.018)
Priming × Flagging	-0.423	-0.423	0.0617	-0.423	-0.423	0.0617
	(0.995)	(0.996)	(1.038)	(0.995)	(0.996)	(1.038)
Mean rating	24.59***	29.52***	29.50***	24.78***	23.84***	23.83***
	(0.989)	(0)	(0)	(0.907)	(0)	(0)
Total number of reviews (N)	0.00393	0.119***	0.119***	0.00291	0.0476***	0.0475***
	(0.00208)	(0)	(0)	(0.00209)	(0)	(0)
Experience good				2.034	3.988***	3.979***
				(1.862)	(1.199)	(1.207)
Priming × Experience good				0.273	0.273	0.275
				(0.350)	(0.350)	(0.356)
Flagging × Experience good				1.196	1.196	1.198
				(2.401)	(2.402)	(2.411)
Dummies for each good	No	Yes	Yes	No	Yes†	Yes†
Demographic and Big Five controls	No	No	Yes	No	No	Yes
Constant	-30.99***	-59.15***	-56.16***	-32.36***	-35.65***	-32.70***
	(3.762)	(0.784)	(2.454)	(3.505)	(0.998)	(2.906)
Observations	7750	7750	7740	7750	7750	7740
R-squared	0.621	0.630	0.634	0.623	0.630	0.634
Adjusted R^2	0.620	0.630	0.633	0.623	0.630	0.633
AIC	65298	65094	64948	65250	65096	64945
BIC	65340	65115	65011	65313	65130	65008

Notes: All regression specifications are estimated using OLS with standard errors clustered by good (i.e. 10 clusters in total). Standard errors are shown in parentheses. The set of demographic variables includes: age, gender, income, employment status, relationship status, and highest level of education. Dummies for goods 9 and 10 were omitted from all regressions for multicollinearity. The AIC and BIC are tests of model fit, where the minimum value suggests the preferred model. † The dummy for good 8 was omitted from (5) and (6) due to multicollinearity. *** $p < .01$, ** $p < .05$.

Treatment effects for individual products across both studies We finally explore differences in treatment effects at a per-product level for additional insight. Figure 3 shows the mean reported quality from the experiment for each good under the main trimming manipulation for each study (T1 for Study 1 and the flagging treatment for Study 2). Unsurprisingly, treatment effects for individual products are highly dependent on the shape of the control rating distribution. T1 in Study 1 and the flagging treatment in Study 2 resulted in exactly the same trimmed distribution, and one can see from Figure 3 that the direction of effect is largely the same for each product. Hence, to some extent, flagging extreme ratings as suspicious does have a similar

effect to removing them entirely. Furthermore, the products with a low valence see an increase in quality, whereas the products with a high valence see a reduction in quality. This suggests that the treatments are helping to create more moderate judgements.

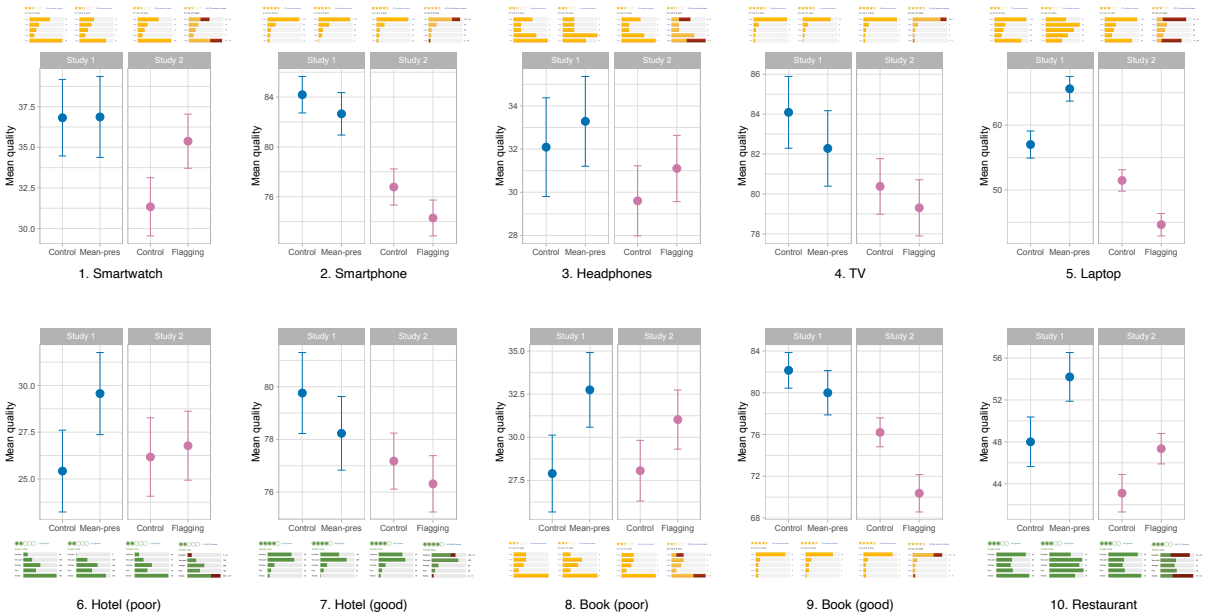


Figure 3: Mean reported quality for each product under the primary rating manipulation for each study. The mean-preserving treatment was used in Study 1, and a review-flagging treatment was used in Study 2. 95% confidence intervals are shown around means. Quality is assessed on a 0-100 scale. The 5 products on the top row are search goods, the 5 on the bottom row are experience goods. Each column corresponds to goods of a similar review distribution e.g. goods 1 and 6 are poorly rated with a high volume of reviews.

Table S4 in the Supplemental Material lists all of the average treatment effects estimated by OLS regressions, without and with additional controls, for each product across both studies. For Study 1, significant treatment effects were found for goods 5, 6, 8, and 10 at least at the 5% level. Large treatment effects are found for goods 5 and 10, but this is because removing extreme ratings essentially inverts the shape of these distributions from a U-shape to a hump-shape. For Study 2, the flagging treatment was significant in changing mean quality ratings for 6 of the 10 goods (3 search and 3 experience). In particular, goods 1, 4, and 9 had significant flagging effects on quality in Study 2 but no significant effects on quality under the mean-preserving treatment in Study 1. It is likely that this was because the reduction of extreme ratings in Study 1 did not visually change the shape of the distribution much for these products. Therefore, in situations where the shape of the distribution is unlikely to change substantially from the removal of fake reviews, flagging suspicious extreme ratings may be more effective in changing judgements than simply removing them entirely. In contrast, the priming treatment in Study 2 only had a

significant treatment effect for a single product. This suggests that informational priming on fake reviews has virtually no impact on judgements.

Model-fitting: Do people apply implicit weights to different ratings?

To test Hypotheses 2a, 2b, and 3b, we must first estimate the weight placed on each possible rating. As explained in the Hypotheses section, the weighted mean is given by:

$$\mu_w = \frac{1}{N}\sum_{r=1}^{5} w_r(rn_r) \tag{3}$$

Therefore, we must estimate the five weights w_r in the following model:

$$Quality = \beta_0 + \beta_1\left[\frac{w_1n_1 + 2w_2n_2 + 3w_3n_3 + 4w_4n_4 + 5w_5n_5}{N}\right] + \epsilon \tag{4}$$

If we let $x_r = \frac{rn_r\beta_1}{N}$ then the regression model can be re-written in the following way:

$$Quality = \beta_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + \epsilon \tag{5}$$

Hence, we can estimate the weights by computing x_1 to x_5. Note that in order to do this, we need to know β_1. However, notice that when $w_r = 1$ for all r, then the weighted mean is simply the mean. In other words, the model becomes:

$$Quality = \beta_0 + \beta_1\mu + \epsilon \tag{6}$$

We can obtain estimates of β_0 and β_1 (i.e. $\hat{\beta}_0$ and $\hat{\beta}_1$) by running a simple regression of quality on mean rating with no additional variables added. Using this estimate allows us to compute $x_r = \frac{rn_r\hat{\beta}_1}{N}$. We can then estimate the weights by using the regression model specified by equation (5), holding β_0 at its estimated level from equation (6), and using the constrained (i.e. restricted) least squares estimator (see, for example, Greene, 2003).

Table 4 shows the estimated weights using constrained least squares regressions. Regression (1) shows that the weights for 1-star, 3-star, and 5-star reviews are all significantly different from 1 when we fit the model on data from Study 1. All observations under T2 were excluded because these observations would have biased the estimated weights. The null hypothesis for each weight is that $w_r = 1$, because if this were true for all $r \in \{1, 2, 3, 4, 5\}$ then no weighting of ratings would be taking place. Therefore, the estimates suggest that extreme ratings are

Table 4: Estimates of the implicit weights applied to ratings of each score

	Dep variable: Reported product quality			
	Study 1		Study 2	
	(1)	(2)	(3)	(4)
x_1	0.203***	-0.0547***	0.341***	0.453***
	(0.124)	(0.156)	(0.0883)	(0.136)
x_2	1.074	1.098	1.242***	1.298**
	(0.145)	(0.131)	(0.0879)	(0.123)
x_3	1.543***	1.669**	1.423***	1.459***
	(0.197)	(0.276)	(0.0948)	(0.0949)
x_4	1.032	1.022	1.150***	1.178***
	(0.0484)	(0.0585)	(0.0222)	(0.0358)
x_5	0.956***	0.937***	0.943***	0.965
	(0.0124)	(0.0157)	(0.0111)	(0.0287)
Includes controls for: treatment, Big Five, and demographics)	No	Yes	No	Yes
Constant (constrained)	-29.84	-29.84	-29.4	-29.4
	(0)	(0)	(0)	(0)
Observations	3341	3341	7750	7740
AIC	26956	26898	65141	64989
BIC	26987	26953	65176	65052

Notes: Regressions were estimated using constrained least squares, where the constant is constrained to the value obtained when fitting a basic model that only includes the mean rating. (1) and (2) are estimated using Study 1 data but exclude the *no-extreme* treatment (T2). (3) and (4) are estimated using Study 2 data. Standard errors are clustered by good, and shown in parentheses. The estimated implicit weights applied to reviews $\hat{w}_r$ are the coefficient estimates for the variables x_r, where $r = 1, 2, 3, 4, 5$ is the review score. Significance for weights is based on the null hypothesis $\hat{w}_r = 1$. *** $p <$.01, ** $p <$ .05. The AIC and BIC are tests of model fit, where a smaller value indicates a better fit.

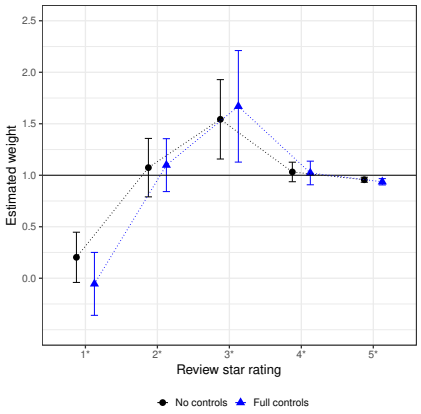
being weighted implicitly by individuals in order to evaluate product quality. Specifically, the estimated weights from (1) imply that a 1-star review is actually valued at approximately 0.203 stars, while a 5-star review is valued at approximately 4.78 stars. Repeating this regression model using data from Study 2 in regression (3) supports the conclusion that 1-star and 5-star reviews are being weighted differently from 1. Additional controls for treatment condition, personality, and demographics are added in regressions (2) and (4) of Table 4 to test the robustness of the weights estimated in (1) and (3). Aside from the fact that the weight on 1-star reviews in regression (2) drops substantially, and the estimated weight on 5-star reviews is no longer statistically significantly different from 1 in regression (4), there is relatively little difference to the overall weighting profile when adding these controls. However, notably both studies found that individuals higher in Openness to Experience significantly rated product quality as being lower. The estimated weights from these four regressions are plotted in Figures 4a and 4b.

If individuals were adjusting for 1-star reviews by assuming that they were fake, or a biased and overcritical evaluation of true quality, one would expect a weight greater than 1 to be

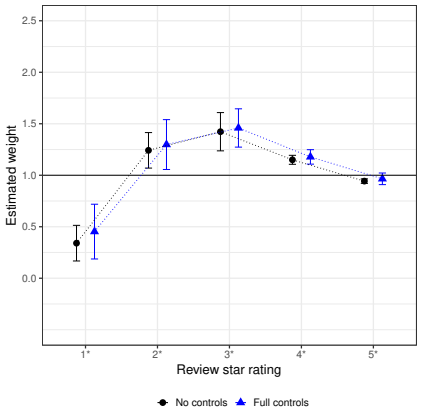
applied. This would moderate extreme ratings towards the middle of the review scale. However, the estimates in Table 4 suggest that 1-star reviews are weighted by less than 0.5. In other words, no moderation process is taking place at the low end of the review scale, and a negativity bias may instead be present, whereby a negative signal about the product is even more harmful than it would be expected to be from a numerical standpoint. In the Hypotheses section, it was explained that a weight on 1-star reviews of 3 and a weight on 5-star reviews of 0.6 would represent an extreme situation where the valence of these reviews was completely eliminated because they would effectively have value equal to an objective 3-star rating. The estimated weights show that neither of these normative extremes is true. For 1-star ratings, there is not even partial adjustment, given that the weights are below 1. For 5-star ratings, the estimated weights in all four regressions are both significantly greater than 0.6 and significantly less than 1. Therefore, individuals seem to be partially adjusting for disingenuous or biased 5-star reviews (Hypothesis 2a holds), but they are not adjusting for disingenuous or biased 1-star reviews (only the alternate version of Hypothesis 2b holds).

In Figure 4c, the average weight estimates from regressions (1) and (3) have been multiplied by their corresponding rating in order to create a mapping between the actual rating and the weighted (i.e. subjectively perceived) rating. From this we can see that 1-star and 5-star ratings are valued at below their objective values, whereas 2, 3, and 4-star ratings are valued above their objective values. To put the estimated weights into perspective, suppose we use the parenting book as an example (Good 9 in Table 1). This has a mean rating of 4.3 stars, and predicted quality of 80.6 (out of 100) using regression (1) in Table 4. Suppose that one new disingenuous 1-star review were added to the distribution, so that the number of 1-star ratings increases from 3 to 4. Perceived quality would fall to 78.4. In order to offset this single 1-star review and revert estimated quality to at least 80.6, a further *eight* 5-star reviews would need to be added. Repeating this exercise with the estimated weights from Study 2 in regression (3), one additional 1-star review would require approximately a further *nine* 5-star reviews to be added in order to offset its impact on rated quality. In sum, while it is true that individuals appear to be applying weights to ratings, and the weight attached to 5-star reviews is less than 1, the strong weight attached to 1-star reviews suggests that negativity bias is likely to override any adjustment of 1-star reviews towards the midpoint of the scale.

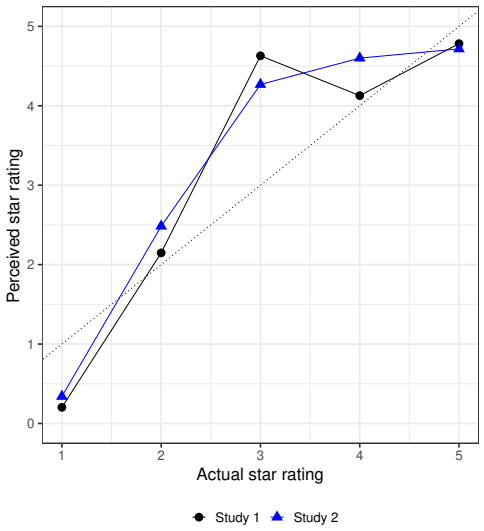
The relationship between type of good and weights. Hypothesis 3b predicts that the weights placed on 1-star and 5-star ratings will be significantly different for experience goods relative to



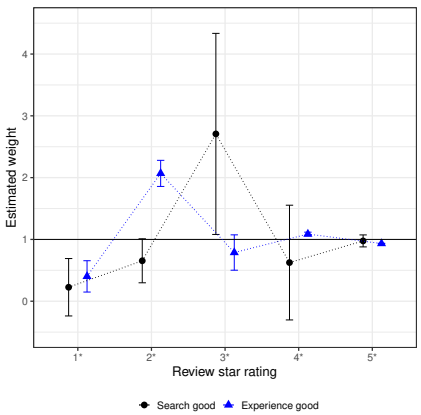
(a) Study 1 excl T2 ($n = 3341$ per regression)



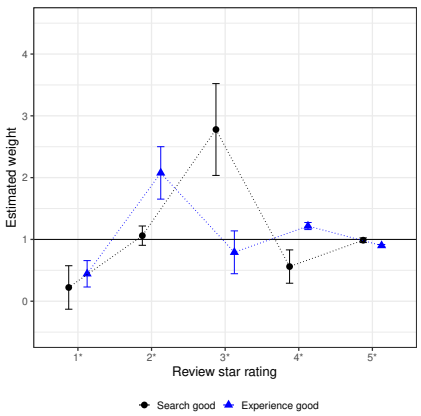
(b) Study 2 ($n = 7750$, 7740 for full controls)



(c) Mapping of actual to perceived review score



(d) Study 1 excl T2 ($n = 1670$ for search, $n = 1671$ for experience)



(e) Study 2 ($n = 3875$ per regression)

Figure 4: (a) and (b): Plots of the estimated implicit weights applied to reviews of a given score, with 95% confidence intervals estimated using standard errors clustered by good. Black circles represent weights estimated from a regression with controls for treatment, demographics, and personality. (c) The mapping between actual and perceived ratings from each study, where the dotted line represents an unweighted 1:1 mapping. (d) and (e): Weights moderated by type of good.

search goods. The regressions in Table 5 suggest that this is not quite true. The first and third regressions in Table 5 fit the weights only to search goods (including and excluding observations under T2), whereas the second and fourth regressions in Table 5 fit the weights only to experience goods. Note that clustering by good was not used for the Study 2 regressions, because this causes standard errors to collapse to zero. No controls were included in these regressions for ease of interpretation of the coefficients, and because the previous analyses suggest that the demographic and personality variables provided very little explanatory power.

The most significant difference in weights between search and experience goods is how 2-star and 3-star ratings are valued. For experience goods, 2-star ratings are weighted by approximately 1.8, which means that their positive impact on perceived quality seems to be higher than 3-star reviews. However, for search goods, 2-star ratings are not weighted at all, whereas 3-star reviews are valued disproportionately highly. This pattern can be seen visually in Figures 4d and 4e. It is not clear why this is the case, though it might reflect the more subjective nature of experience goods relative to search goods. The high weight on 2-star ratings could indicate that individuals are more aware that for experiences, a slightly poor review is more likely to reflect a reviewer's preferences rather than indicating that the experience is genuinely of poor quality. This means that only 1-star ratings are likely to be viewed as a negative signal of quality for experience goods, but 1-star and 2-star ratings are both likely to be viewed as a negative signal of quality for search goods. Materially, however, this difference in weighting profile is unlikely to make much difference in terms of the overall impact of fake reviews. Returning to the example from Good 9, the addition of a 1-star review to the distribution reduces perceived product quality by 1.9 for both search and experience goods when using the models estimated from Study 1 data, and 2 and 1.9 for search and experience goods respectively when using the models estimated from Study 2 data. Therefore, Hypothesis 3b does not hold - the weight on extreme ratings is not different between search goods and experience goods. However, the results suggest that there may be differences in the perception of *moderate* reviews for experience goods, relative to search goods.

Exploratory analysis: the moderating effect of priming and flagging on weights. Finally, we consider whether the priming and the flagging treatments in Study 2 affect estimated weights. Neither of these treatments change the underlying ratings distribution, but both provide an indication that some of the reviews may be biased or fake, which might suggest a change in weighting at the extremes. Testing the moderating effect of these treatments allows us to de-

Table 5: Moderating effect of good type and review treatments on implicit weights attached to ratings

	Dependent variable: Reported product quality							
	Study 1 (no T2)		Study 2					
	Search	Experience	Search	Experience	No-flagging	Flagging	No-priming	Priming
x_1	0.226***	0.402***	0.222***	0.444***	0.308***	0.377***	0.349***	0.332***
	(0.237)	(0.129)	(0.179)	(0.109)	(0.179)	(0.112)	(0.101)	(0.0858)
x_2	0.654*	2.069***	1.061	2.076***	1.225**	1.261*	1.249***	1.235***
	(0.182)	(0.108)	(0.0797)	(0.216)	(0.113)	(0.146)	(0.114)	(0.0656)
x_3	2.708**	0.788	2.779***	0.791	1.525***	1.309**	1.407***	1.439***
	(0.830)	(0.146)	(0.379)	(0.177)	(0.0926)	(0.141)	(0.109)	(0.0834)
x_4	0.625	1.087***	0.561***	1.217***	1.093***	1.213***	1.152***	1.148***
	(0.473)	(0.0155)	(0.138)	(0.0291)	(0.0192)	(0.0403)	(0.0249)	(0.0200)
x_5	0.976	0.933***	0.987	0.901***	0.951***	0.935***	0.943***	0.943***
	(0.0495)	(0.00203)	(0.0194)	(0.00880)	(0.0103)	(0.0206)	(0.0119)	(0.0104)
Cons	-28.28	-31.8	-31.86	-28.85	-33.33	-25.44	-29.69	-29.09
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
Obs	1670	1671	3875	3875	3890	3860	3960	3790
AIC	13445	13430	32415	32671	32799	32309	33215	31934
BIC	13466	13452	32415	32671	32831	32340	33246	31965

Notes: Significance for weight estimates is based on the null hypothesis $\hat{w}_r = 1$. Estimated weights w_r are the coefficient estimates of the variables x_r, where $r \in \{1, 2, 3, 4, 5\}$ is the review score. The weight for 3-star reviews is constrained to 1, and the constant is constrained to the value obtained by a simple univariate regression on the mean review score. The first two regressions utilise the full sample; the final two exclude observations in T2, where no 1-star or 5-star reviews were shown. Standard errors are clustered by good (apart from for the search and experience regressions in Study 2), and shown in parentheses. *** $p < .01$, ** $p < .05$, * $p < .1$. The AIC and BIC are tests of model fit, where the minimum value suggests the preferred model.

termine whether informational interventions about fake reviews are actually likely to change how people judge product quality from rating distributions. The final four columns in Table 5 show the estimated weights based on treatment type, and Figure S2 in the Supplemental Material plots these graphically. Flagging appears to lead to a slightly flatter weighting profile than for non-flagged distributions, with 1-star ratings being down-weighted slightly less, and 5-star ratings being down-weighted slightly more. This suggests that flagging may help to mitigate for extreme rating bias, although the differences in extreme rating weights are not substantially different to the no-flagging case. Returning to the ratings distribution used in Good 9, the estimates suggest that predicted quality using the weighted model = 74.3 without flagging, and 71.0 with flagging. Therefore, flagging suspicious extreme reviews may slightly offset the impact of fake / biased bad reviews, but it is unlikely to change perceptions enough to affect actual choice. Priming people to the possibility of fake reviews appears to have virtually no effect at all in changing how they weight ratings of different scores. Continuing the previous example with Good 9, there is only a 0.4 point difference in predicted quality when using the model estimated without priming vs with priming. Hence, priming in the context of fake reviews does

not affect the judgement process.

Discussion

This study has investigated how individuals judge product quality based on numerical rating distributions, whether they adjust for potential bias in extreme ratings, and whether trimming extreme ratings can influence judgements. The study has two main findings. First, trimming 1-star and 5-star ratings from a distribution, akin to an externally enforced version of the trimmed mean heuristic (Yaniv, 1997), may change judgements of product quality when the mean rating is preserved. However, this effect is not statistically significant when treatment effects are estimated using regressions with clustered standard errors. This implies that on average, the mean rating overrides small changes in the ratings distribution. If the trimming is performed by only flagging (rather than removing) the outlying ratings, then there is no significant effect on quality judgements. However, the direction of the treatment effects on individual products is generally the same whether we flag extreme ratings or remove them entirely. Externally enforced trimming or flagging can reduce the extremity of judgements for particular product-distribution combinations, but appears to have little impact on aggregate. Flagging appears particularly effective in situations where removing extreme ratings completely would not have a perceptible impact on the shape of the overall distribution. Whether the product in question was a search good or an experience good does not moderate the effectiveness of these trimming manipulations. Furthermore, priming individuals to the possibility that reviews may be fake or biased does not affect judged quality at all.

Together, these results suggest that: (i) explicitly making consumers aware of biased ratings does not change judgements and is therefore unlikely to be an effective intervention to improve decision-making; (ii) trimming extreme ratings by removing them entirely may be more effective, but only if it changes the mean rating; and (iii) flagging suspicious extreme ratings can be preferable to removal in terms of changing judgements in situations where removing ratings does not significantly change the ratings distribution. In sum, the mean rating is difficult to override as the main signal of judged product quality, and decision aids designed to help a consumer adjust for disingenuous reviews would need to change the mean to change judgements sufficiently.

Second, individuals implicitly weight ratings in an asymmetric hump-shaped fashion, where 1-star ratings are weighted more heavily than 5-star ratings. This weighting scheme is relatively

stable across both studies, and is not responsive to indicators of fake or biased reviews, though it may change slightly based on the type of product being evaluated. The estimated weights suggest that product quality judgements are more likely to be swayed by disingenuous 1-star reviews than disingenuous 5-star reviews. People appear to be slightly down-weighting 5-star reviews, so that they are valued similarly to 3-star and 4-star reviews. One explanation for why people may be adjusting for bias in 5-star ratings but not bias in 1-star ratings may be consumer desensitisation towards extremely positive reviews, because they are seen as the norm (e.g. Chen & Lurie, 2013). Consumers are likely to be used to seeing a high volume of 5-star ratings because the majority of ratings distributions are J-shaped (Hu et al., 2009). A consumer may reasonably be assuming that a majority of products cannot be above average in terms of objective quality, and therefore the relative value of a 5-star rating needs to be reduced in order to correct for this bias. However, they are not up-weighting 1-star reviews in order to compensate for the potential impact of biased bad reviews, perhaps because these are less common. This means that even a small number of disingenuous 1-star reviews can have a large adverse impact on perceived product quality. A proportionally large number of positively-valenced reviews would be required in order to offset the impact of a small number of 1-star reviews. The precise ratio of 5-star ratings required to offset a 1-star rating depends on the initial ratings distribution.

The estimated weighting profiles can be linked to two main psychological phenomena. First, the strong weighting of 1-star ratings is consistent with negativity bias (Baumeister et al., 2001). The emergent rule-of-thumb from prior research is that there needs to be at least five good events in order to offset each bad event (Baumeister et al., 2001). Estimates from the present studies suggest that the ratio for the number of 5-star reviews needed to offset each 1-star review may be even greater than 5:1 for rating distributions with a typical J-shape. For example, using one of the sample rating distributions from the study yielded ratios of 8:1 and 9:1. Second, when the weights are used to map objective rating to subjective rating, 1-star reviews contribute less than 1-star to the subjective mean rating, 2-star reviews contribute approximately 2-stars to the subjective mean rating, and 3-star to 5-star reviews contribute between 4-stars and 5-stars to the subjective mean rating. This is reminiscent of a binary bias in that ratings appear to be placed into discrete bins (Fisher & Keil, 2018; Fisher et al., 2018), although in our case there appear to be three bins rather than two. The result may still indicate a binary bias, but with an inflection point at 2-stars, as opposed to Fisher et al. (2018)'s suggestion that 3-star ratings are

the inflection point. In other words, our weights suggest that 2-star ratings are interpreted as a neutral signal of quality, with only 1-star ratings providing a negative signal, and anything above 2-stars providing a positive signal.

Finally, moderate ratings (but not extreme ratings) appear to be weighted slightly differently depending on whether the product is a search good or an experience good. A much larger (positive) weight is placed on 2-star ratings for experience goods than for search goods, and a much larger (positive) weight is placed on 3-star ratings for search goods than experience goods. This finding appears to reflect an understanding in consumers that experience goods (i.e. horizontally-differentiated products) are more dependent on taste than search goods (i.e. vertically-differentiated products), and so have less of a clear objective valuation of quality (e.g. H. Dai, Chan, & Mogilner, 2020). Knowing this, consumers may overvalue 2-star ratings because they perceive them as expressing differing tastes rather than indicating low quality. However, 1-star ratings for experience goods are weighted in much the same way as they are for search goods.

One limitation of this study is that we cannot say anything about how quantitative ratings interact with textual information in judgement formation. Written reviews are undoubtedly important in determining the usefulness of a review in any individual's belief updating process (e.g. Hu, Liu, & Zhang, 2008; Ludwig et al., 2013; Mudambi & Schuff, 2010; Mudambi, Schuff, & Zhang, 2014), as are other elements such as reviewer identity (Forman, Ghose, & Wiesenfeld, 2008). However, as well as being important for theoretical understanding, isolating the quantitative component of reviews may still be relevant in real-world situations. Consumers may not have time to read through the text of a large number of reviews, or may not wish to exert the effort required to do so, especially when the number of products under consideration is high. Some websites also only allow for numerical ratings. Other than this, the studies omitted brand information to prevent participants from using previously formed beliefs, but this may have exacerbated negativity bias due to unfamiliarity (e.g. Ahluwalia, 2002).

Another possible limitation with the design of Study 1 is that individuals saw a different treatment distribution for each product, which in theory could lead to carryover effects. However, this is unlikely to be the case for two main reasons: first because the order of products was randomised for each participant, and second because it would not have been obvious that they were seeing treated distributions at all. The between-subjects design of Study 2 is generally unlikely to suffer from such issues, and the general agreement between the two studies in terms

of estimated judgements and weights adds robustness and confidence to the findings.

An important next step for this line of research would be to develop a process model in order to understand how the weighting profile found in this study arises, both from cognitive reasoning as well as perception. Any approach would need to be able to sufficiently explain the evidence from the present studies, i.e. a strong focus on the mean rating, heavier weighting at the extreme ends of the scale with a negativity bias, and a binary bias with an inflection point at 2-stars. One approach could be to assume that weights are derived from multiple higher-level social factors, such as trust. For example, consider:

$$w_r = f_r(\text{product attributes}) t_r(n_r, ...) \tag{7}$$

Here, $f_r(.)$ represents a function that determines a weight for rating r based on objective product attributes. For example, if the product was a search good, then it is likely from our findings that $f_3(.) > 1$. $t_r(.)$ represents a "trust" function that takes values between 0 and 1 that would augment this initial weight. Trust may increase as the volume of ratings of score r increases, but may also be dependent on other factors such as prior knowledge of a particular marketplace or suspected manipulation. If trust for reviews of rating r falls $t_r(.)$ will fall, bringing the weight closer to 0. Another approach would be to combine existing psychological models to offer an explanation based on information processing rather than deductive reasoning. For example, one could start with a range-frequency approach (e.g. Tripp & Brown, 2016) where the mean rating and other product cues form a signal that should be interpreted relative to its position within the distribution of ratings, perhaps also comparing this to an "expected" ratings distribution from memory. Binary bias and negativity bias could then be worked into the model design.

Concluding remarks

This study has shown that individuals implicitly apply weights to different ratings, suggesting that they may be aware (to some degree) that a rating is not an unbiased signal of product quality. Despite this, trimming extreme ratings without changing the mean does not affect judgements on average, though it may do for particular rating distributions. This is true even when information about potential review bias is made explicit. Therefore, consumers are heavily reliant on the mean rating to judge product quality, at the expense of other distributional signals. They are particularly susceptible to the presence of disingenuous 1-star reviews, because the

weight placed on 1-star ratings exacerbates bias rather than compensates for it. Disingenuous 1-star reviews have a high risk of adversely impacting consumer judgement, leading to suboptimal choices. Future research efforts should be directed towards (1) understanding *how* individuals form weights for ratings, and (2) testing mechanisms that adjust the mean rating or reduce its salience.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When Should You Adjust Standard Errors for Clustering? *Quarterly Journal of Economics*, *138*(1), 1–35. doi:10.1093/qje/qjac038

Ahluwalia, R. (2002). How Prevalent Is the Negativity Effect in Consumer Environments? *Journal of Consumer Research*, *29*(2), 270–279. doi:10.1086/341576. eprint: https://academic.oup.com/jcr/article-pdf/29/2/270/5207258/29-2-270.pdf

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science*, *12*(2), 157–162. doi:10.1111/1467-9280.00327

Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly*, *26*(3), 243–268. Retrieved from http://www.jstor.org/stable/4132332

Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, *53*(3), 297–318. doi:10.1509/jmr.14.0380

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

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. doi:10.1016/j.obhdp.2006.07.001

Box, D., & Croker, S. (2018). *Fake five-star reviews being bought and sold online*. Retrieved from https://www.bbc.co.uk/news/technology-43907695

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science*, *22*(3), 384–392. doi:10.1177/0956797610397956

Brodkin, J. (2021). *Posing as amazon seller, consumer group investigates fake-review industry*. Retrieved from https://arstechnica.com/tech-policy/2021/02/fake-amazon-reviews-still-sold-in-bulk-it-costs-10900-for-1000-reviews/

Brown, G. D. A., & Matthews, W. J. (2011). Decision by Sampling and Memory Distinctiveness: Range Effects from Rank-Based Models of Judgment and Choice. *Frontiers in Psychology*, *2*, 13972. doi:10.3389/fpsyg.2011.00299

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. doi:10.3368/jhr.50.2.317

Chen, Z., & Lurie, N. H. (2013). Temporal contiguity and negativity bias in the impact of online word of mouth. *Journal of Marketing Research*, *50*(4), 463–476. doi:10.1509/jmr.12.0063

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. doi: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. doi:10.1287/mksc.1100.0572

Dai, H., Chan, C., & Mogilner, C. (2020). People Rely Less on Consumer Reviews for Experiential than Material Purchases. *Journal of Consumer Research*, *46*(6), 1052–1075. doi:10.1093/jcr/ucz042. eprint: https://academic.oup.com/jcr/article-pdf/46/6/1052/33038486/ucz042.pdf

Dai, W., Jin, G., Lee, J., & Luca, M. (2018). Aggregation of consumer ratings: An application to yelp.com. *Quantitative Marketing and Economics*, *16*(3), 289–339. doi:https://doi.org/10.1007/s11129-017-9194-9

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

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-ipip scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, *18*(2), 192. doi:10.1037/1040-3590.18.2.192

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. doi:10.1002/agr.21653

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993. doi:10.1037/xge0000098

Fisher, M., & Keil, F. C. (2018). The Binary Bias: A Systematic Distortion in the Integration of Information. *Psychological Science*, *29*(11), 1846–1858. doi: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*, *45*(3), 471–489. doi:10.1093/jcr/ucy017

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291–313. doi:10.1287/isre.1080.0193

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gilovich, T., & Gallo, I. (2020). Consumers' pursuit of material and experiential purchases: A review. *Consumer Psychology Review*, *3*(1), 20–33. doi:10.1002/arcp.1053

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. doi:10.1016/j.jrp.2005.08.007

Greene, W. H. (2003). *Econometric analysis* (5th edition). Upper Saddle River, New Jersey: Prentice Hall.

Guha Majumder, M., Dutta Gupta, S., & Paul, J. (2022). Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis. *Journal of Business Research*, *150*, 147–164. doi:10.1016/j.jbusres.2022.06.012

Harries, C., Yaniv, I., & Harvey, N. (2004). Combining advice: the weight of a dissenting opinion in the consensus. *Journal of Behavioral Decision Making*, *17*(5), 333–348. doi:10.1002/bdm.474

He, S., Hollenbeck, B., & Proserpio, D. (2022). The Market for Fake Reviews. *Marketing Science*. Retrieved from https://pubsonline.informs.org/doi/abs/10.1287/mksc.2022.1353

Heiniger, S., & Mercier, H. (2018). National Bias of International Gymnastics Judges during the 2013-2016 Olympic Cycle. *ArXiv e-prints*. doi:10.48550/arXiv.1807.10033. eprint: 1807.10033

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, *52*(3), 674–684. doi:10.1016/j.dss.2011.11.002

Hu, N., Liu, L., & Sambamurthy, V. (2011). Fraud detection in online consumer reviews. *Decision Support Systems*, *50*(3), 614–626. On quantitative methods for detection of financial fraud. doi:https://doi.org/10.1016/j.dss.2010.08.012

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? the role of reviewer characteristics and temporal effects. *Information Technology and Management*, *9*(3), 201–214. doi:10.1007/s10799-008-0041-2

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. doi: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. doi:10.1145/1562764.1562800

Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education*, *84*(1), 175–196. doi:https://doi.org/10.1080/00220973.2014.952397

Huang, P., Lurie, N. H., & Mitra, S. (2009). Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods. *Journal of Marketing*, *73*(2), 55–69. doi:10.1509/jmkg.73.2.55

King, R. A., Racherla, P., & Bush, V. D. (2014). What we know and don't know about online word-of-mouth: A review and synthesis of the literature. *Journal of Interactive Marketing*, *28*(3), 167–183. doi:https://doi.org/10.1016/j.intmar.2014.02.001

Klein, L. R. (1998). Evaluating the potential of interactive media through a new lens: Search versus experience goods. *Journal of Business Research*, *41*(3), 195–203. doi:https://doi.org/10.1016/S0148-2963(97)00062-3

Köcher, S. [Sarah], & Köcher, S. [Sören]. (2021). The Mode Heuristic in Service Consumers' Interpretations of Online Rating Distributions. *Journal of Service Research*, *24*(4), 582–600. doi:10.1177/10946705211012475

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. doi:10.1016/j.elerap.2010.04.001

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. doi: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. doi:10.1016/j.chb.2020.106670

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., & Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. In *Proceedings of the 19th acm international conference on information and knowledge management* (pp. 939–948).

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). doi:10.1007/978-1-4614-3223-4_13

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp.com. *Harvard Business School Working Paper*, (12-016). doi:10.2139/ssrn.1928601

Ludvig, E. A., Madan, C. R., McMillan, N., Xu, Y., & Spetch, M. L. (2018). Living near the edge: How extreme outcomes and their neighbors drive risky choice. *Journal of Experimental Psychology. General*, *147*(12), 1905–1918. doi:10.1037/xge0000414. eprint: 29565605

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, *77*(1), 87–103. doi:https://doi.org/10.1509/jm.11.0560

Lyon, A., Wintle, B. C., & Burgman, M. (2015). Collective wisdom: Methods of confidence interval aggregation. *Journal of Business Research*, *68*(8), 1759–1767. doi:10.1016/j.jbusres.2014.08.012

Marinescu, I., Chamberlain, A., Smart, M., & Klein, N. (2021). Incentives can reduce bias in online employer reviews. *Journal of Experimental Psychology: Applied*, *27*(2), 393.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, *104*(8), 2421–2455. doi:10.1257/aer.104.8.2421

McAuley, J., Pandey, R., & Leskovec, J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM.

McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 43–52). ACM.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? a study of customer reviews on amazon.com. *MIS Quarterly*, *34*(1), 185–200. doi:10.2307/20721420

Mudambi, S. M., Schuff, D., & Zhang, Z. (2014). Why aren't the stars aligned? an analysis of online review content and star ratings. *47th Hawaii International Conference on System Science*, 3139–3147. doi:10.1109/HICSS.2014.389

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on world wide web* (pp. 191–200). ACM. doi:10.1145/2187836.2187863

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311–329. doi:https://doi.org/10.1086/259630

Nelson, P. (1974). Advertising as information. *Journal of Political Economy*, *82*(4), 729–754. doi:https://doi.org/10.1086/260231

Ocean, N. (2023). Weighting extreme ratings. doi:https://doi.org/10.17605/OSF.IO/9ZKWN

Parducci, A. (1968). The relativism of absolute judgments. *Scientific American*.

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

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European review of social psychology*, *1*(1), 33–60. doi:10.1080/14792779108401856

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865. doi:10.1037/xge0000465

Putnam-Farr, E., & Morewedge, C. K. (2021). Which social comparisons influence happiness with unequal pay? *Journal of Experimental Psychology: General*, *150*(3), 570. doi:http://dx.doi.org/10.1037/xge0000965

R Core Team. (2023). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. doi:10.1207/S15327957PSPR0504_2

Song, W., Park, S., & Ryu, D. (2017). Information quality of online reviews in the presence of potentially fake reviews. *Korean Economic Review*, *33*(1), 5–34.

Spiller, S. A., & Belogolova, L. (2017). On Consumer Beliefs about Quality and Taste. *Journal of Consumer Research*, *43*(6), 970–991. doi:10.1093/jcr/ucw065

StataCorp. (2013). Stata statistical software: Release 13. *College Station, TX: StataCorp LP*.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. doi: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. doi:10.1287/mnsc.1110.1458

Sun, X., Han, M., & Feng, J. (2019). Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products. *Decision Support Systems*, *124*, 113099. doi:10.1016/j.dss.2019.113099

Sunder, S., Kim, K. H., & Yorkston, E. A. (2019). What Drives Herding Behavior in Online Ratings? The Role of Rater Experience, Product Portfolio, and Diverging Opinions. *Journal of Marketing*, *83*(6), 93–112. doi:10.1177/0022242919875688

Tripp, J., & Brown, G. D. A. (2016). Being paid relatively well most of the time: Negatively skewed payments are more satisfying. *Memory and Cognition*, *44*(6), 966–973. doi:10.3758/s13421-016-0604-0

U.S. Department of Commerce. (2022). *Quarterly retail e-commerce sales*. Retrieved from https://www.census.gov/retail/index.html

Wu, Y., Ngai, E. W., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems*, *132*, 113280. doi:https://doi.org/10.1016/j.dss.2020.113280

Yaniv, I. (1997). Weighting and Trimming: Heuristics for Aggregating Judgments under Uncertainty. *Organizational Behavior and Human Decision Processes*, *69*(3), 237–249. doi:10.1006/obhd.1997.2685

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120. doi:10.1016/j.obhdp.2006.05.006