Consumers and Managers Reject (Superior) Algorithms Because They Fail to Compare Them to the (Inferior) Alternative

Berkeley J. Dietvorst, University of Chicago, USA

In five experiments, I find that consumers and managers often choose (inferior) human judgment over (superior) algorithms (e.g. recommender systems) because they fail to compare algorithms’ performance to that of human judgment. Instead they decide whether or not to use an algorithm by comparing its performance to their performance goal.

[to cite]:


[url]:

http://www.acrwebsite.org/volumes/1024102/volumes/v45/NA-45

[copyright notice]:

This work is copyrighted by The Association for Consumer Research. For permission to copy or use this work in whole or in part, please contact the Copyright Clearance Center at http://www.copyright.com/.
Understanding the Use of Online Reviews and Recommendations in Consumer Judgment and Decision-Making

Chair: Elizabeth C. Webb, Columbia University, USA

Paper #1: The Drivers and Downstream Consequences of the J-Shaped Distribution of Consumer Online Reviews
Verena Schoenmueller, Columbia University, USA
Oded Netzer, Columbia University, USA
Florian Stahl, University of Mannheim, Germany

Paper #2: Using Reviews to Determine Preferences: How Variance in Customer-Generated Reviews Affects Choice
Elizabeth C. Webb, Columbia University, USA
Itamar Simonson, Stanford University, USA

Paper #3: “Don’t Tell Me What to Do!” Shoppers Rely Less on Consumer Reviews for Experiential than Material Purchases
Hengchen Dai, Washington University in St. Louis, USA
Cindy Chan, University of Toronto, Canada
Cassie Mogilner, UCLA, USA

Paper #4: Consumers and Managers Reject (Superior) Algorithms Because They Fail to Compare Them to the ( Inferior) Alternative
Berkeley J. Dietvorst, University of Chicago, USA

SESSION OVERVIEW

Product reviews and recommendations have become a ubiquitous facet of the consumer landscape. For this reason, it is important to understand how and when consumers use online reviews and recommendations. Specifically, the research in this session strives to understand the general features of online reviews (Paper 1), how the characteristics of online reviews, such as their variance, can affect consumer preferences (Paper 2), how consumers differentially use online reviews for experiential and material goods (Paper 3), and how consumers choose between user-generated recommendations and algorithmic recommendation systems (Paper 4).

In Paper 1, Schoenmueller et al. investigate the underlying causes of the positive skew (J-shaped distribution) found in many online reviews. The authors analyze over 130 million reviews to show that the most prevalent distribution of reviews is indeed J-shaped (heavily skewed towards the high end of the review scale). While there is some platform dependency in the shape of the review distributions, the authors show that the main driver is one of self-selection. Specifically, consumers are more likely to post reviews for products that they have evaluated extremely. This self-selection may lead to loss of information and bias in consumers’ reliance on average review ratings as a source of information.

While Paper 1 evaluates the underlying causes of review distributions, Paper 2 investigates how a key feature of review distributions—variance—affects consumer choice. Specifically, Webb and Simonson show that the variance in user-generated reviews, holding average rating constant, moderates various context effects. The authors find that when consumers evaluate products with high variance (versus low variance) in the reviews, they choose less variety, they show a reduced susceptibility to the compromise effect, and they are less likely to defer choice. The authors find that these effects occur because consumers find low variance reviews less useful, leading them to rely on relative value considerations or risk-reduction strategies instead.

In comparison to Paper 2, which focuses on a feature of reviews that moderates reliance on reviews (variance), Paper 3 identifies another moderator—product type. Dai et al. find that consumers rely on reviews less when they are for experiential (vs. material) goods. The authors find this occurs because consumers believe their preferences are more unique for such product types and are thus not well-represented by the reviews of other customers. This underlying process—perceived preference uniqueness—acts as a moderator for the use of consumer reviews in decision-making.

Finally, in Paper 4, Dietvorst investigates why consumers and managers often prefer human generated recommendations and judgments over those generated by algorithms. Dietvorst finds that consumers and managers fail to use algorithms that outperform human judgment because they hold algorithms to a higher standard. Instead of choosing algorithms that beat the best alternative (human judgment), many people only choose algorithms that also meet their lofty goals for forecasting performance.

Together, the research in this session helps consumer researchers and marketers better understand the psychological processes underlying the use of online reviews and recommendations in consumer decision-making and judgment.

The Drivers and Downstream Consequences of the J-Shaped Distribution of Consumer Online Reviews

EXTENDED ABSTRACT

Consumer online reviews have become an integral part of consumers’ decision-making process. Based on a recent study, 90 percent of consumers read online reviews before visiting a store (Invespro 2015) and 88 percent of consumers trust online reviews as much as a personal recommendation (DeMers 2015). Consumer reviews have also been shown to have an economic impact (e.g., Chevalier and Mayzlin 2006; Liu 2006; Moe and Trusov 2012). Recently, several meta-analyses summarize the impact of the number (volume) as well as the average rating (valence) of reviews on products sales (Babić et al. 2016; Floyd et al. 2014; You, Gautham, and Joshi 2015).

One common finding in the study of online reviews has been that reviews are heavily skewed to the positive end of the rating scale, with a few reviews in the mid-range and some reviews at the negative end of the scale (Hu et al. 2009). This has been often referred to as the J-shaped distribution of online ratings (e.g., Dalvi, Kumar, and Pang 2015; Feng et al. 2012; Gao et al. 2015; Hu et al., 2009). This finding is surprising given that online reviews represent crowdsourcing of experiences of a large body of heterogeneous consumers, which based on the law of large numbers and various findings should converge towards a normal distribution. The extreme positive skewness of reviews has fueled the debate on how informative consumer reviews actually are (Fritz 2016; Hickey 2015), and whether these consumer reviews mirror “true” product quality (De Langhe, Fernbach, and Lichtenstein 2015).

While several studies have demonstrated the existence of the skewness of rating distributions, the prevalence, the reasons behind, and the implications of the J-shaped distribution of online reviews has been given considerably less attention compared to the impact of overall rating statistics such as the volume and valence on product sales. Accordingly, the objective of this study is threefold. First, we test how prevalent and robust the J-shaped distribution of reviews is by collecting a large-scale data set across various online review and e-commerce platforms that span across a large number of products.
and services. Second, we investigate what could be the main reasons for the J-shaped distribution of reviews and to what extent it reflects a bias or self-selection of consumers’ preferences and opinions. We employ a multi-pronged approach, including: secondary data analysis, experiments, and surveys. Third, we investigate the information loss and the bias introduced by the J-shaped distribution in capturing consumer preferences and behavior.

Using a large-scale dataset of over 130 million online reviews from 18 different platforms such as Amazon, Yelp, and Expedia, involving more than 10 million products, reflecting different types of platforms (e.g., e-commerce sites, reviews sites, comparison sites) various product categories (e.g., books, beers, hotels, restaurants and services) and various formats of review scales, our analysis reveals that, while indeed the most dominant distribution of reviews across platforms and product categories is a J-shaped distribution, several platforms have consumer reviews and ratings that are not J-shaped distributed and the distribution of the ratings of the same product or service can differ across platforms.

Our results show that while some product selection by the platforms exists, the main driver behind the J-shaped distribution is a reviewer level self-selection mechanism. Using secondary datasets, survey data as well as data collected in several lab experiments, we find that consistent with consumers’ tendency to report and spread WOM for extreme experiences (Anderson 1998), consumers are more likely to review products or services for which the experience or evaluation was extremely positive or extremely negative. We name this polarity self-selection. Specifically, we find that reviewers who are less selective in the products they chose to review exhibit less polarity in their reviews. Furthermore, consumers who were asked to review the last book they read or restaurant they visited, gave less extreme ratings than consumers who selected the product they wish to review. Furthermore, we disentangle the effect of polarity self-selection from purchasing self-selection.

Building on empirical evidence from our large-scale dataset of reviews that suggests that shorter scales may lead to scale truncation as a potential driver of the J-shaped distribution, we use several experiments to test the impact of rating scales on review distributions. Our experiments rule out that the J-shaped distribution is driven by factors related to the usage of different rating scales or wording across online platforms. We further find evidence that cognitive dissonance increases the positive skewness of consumer ratings. Finally, we demonstrate that review fraud possibly explains ratings at the negative extreme of the rating scale but not at the positive extreme of the J-shaped distribution.

We find that the J-shaped distribution of reviews, while somewhat decreasing over time, is fairly robust over the product and the reviewer lifecycle. In summary, our results show that the common J-shaped distribution of online reviews mainly emerges from two selection mechanisms and that reviewers’ self-selection has an impact over and beyond platforms’ product selection.

Finally, we find that polarity self-selection, causing the J-shape distribution, leads to an information loss. We show that for review distributions with a high intensity of self-selection, the average ratings metric commonly used in online review platforms and academic research is less related to the actual sales than the number of reviews. Additionally, we find that inconclusive results in previous research regarding the relationship between the average ratings and sales can be explained by polarity self-selection and the predominance of the J-shaped distribution.

Using Reviews to Determine Preferences: How Variance in Customer-Generated Reviews Affects Choice

EXTENDED ABSTRACT

User-generated product reviews have become an important input in consumer decision-making (Floyd et al. 2014; de Langhe, Fernbach, and Lichtenstein 2016; Simonson 2016). Consumers are now able to use product reviews to learn about products and gather information about their options. While past research has found preferences can be manipulated by context (e.g., the compromise effect, variety-seeking, and choice deferral) (Simonson 1989; 1990; Simonson and Tversky 1992; Kahn 1995; Dhar 1997) the question of how product review information affects these choice effects remains. Beyond the sheer content of reviews, their observed variance and the implied consensus (or lack thereof) among reviewers provides additional, potentially useful information about the drivers of product preferences and the expected consumer experience.

Thus, in this research we examine the impact of user-generated review variance or dispersion, and especially, the implications for consumer susceptibility to context effects. We show that product reviews with greater variance/Dispersion attenuate variety-seeking, reduce choice shares for a compromise option, and decrease choice deferral rates relative to products associated with reviews with less variance/Dispersion. Greater review variance indicates that evaluations are a matter of taste (Spiller and Belogolova 2016), thereby allowing the consumer to deviate from the safe norms (e.g., compromising or hedging). Paradoxically, this means that dispersed reviews are more useful and informative and encourage consumers to follow their own tastes based on the options’ absolute values.

In Study 1 (N = 279), we tested the effect of review variance on variety seeking. Participants were assigned to one of two between-subjects conditions (High vs. Low review variance). Participants in both conditions were told to imagine they were grocery shopping through an online grocer and were asked to choose options across several product categories (yogurt, ice cream, soup, and snacks) for three weeks of consumption. Participants viewed six options in each category. They were informed that each product had reviews from customers who had purchased the product in the particular store. Participants in the High condition viewed options that had high variance/widely distributed product reviews (reviews that were dispersed across the entire rating scale); participants in the Low condition viewed options that had low variance/narrowly distributed product reviews (reviews that were clustered at the high-end of the rating scale). All products in both conditions and across categories had the same average rating (4/5 stars), and only the variance differed across options. We found that for all product categories, participants who chose among options with high variance reviews chose significantly less variety than participants who chose from options with low variance reviews (p < 0.04). We also found that across conditions, participants believed they would feel significantly happier (p = 0.03) and marginally significantly more personal responsibility (p < 0.10) for a positive outcome with a product associated with higher variance reviews. This finding is consistent with the notion that higher review variability allows consumers to feel that they “own” the choice based on their personal tastes.

In Study 2 (N = 393), we investigated whether review variance moderates the compromise effect. In this study, we used a standard compromise effect (two set) test with participants assigned to either a 2-option or 3-option choice set; we compared the relative shares for the middle option between the two sets. Participants made choices in two product categories: TVs and restaurants. The design was a 2 (Variance: High, Low) x 2 (2-option, 3-option), such that partici-
pants saw high variance reviews for one product and low variance reviews for the other (and 2-options for one and 3-options for the other). Again, the reviews all had the same average rating (4/5 stars), but the variance differed across products and was either high or low (depending on condition assignment). We found that choice shares for the compromise (middle) option were significantly higher when participants chose from products with low variance reviews ($p < 0.001$ for TVs; $p = 0.03$ for restaurants). Thus, consumers are less likely to choose a compromise option when they are choosing across products with high variance reviews relative to products with low variance reviews.

In Study 3 ($N = 425$), we evaluated how ratings variance affected choice deferral. Participants were assigned to one of two between-subjects conditions: High variance or Low variance. Using a standard choice deferral set-up, participants were shown two product options from a randomly assigned product category (Microwave or Coffee Maker). Participants in the High (Low) condition, chose between two options with high (low) variance customer reviews. Participants also had the option not to choose (deferr). We found that deferral rates were marginally significantly lower across both product types when participants were in the High condition relative to the Low condition ($p = 0.08$). This suggests that consumers are less likely to defer choice when they are choosing across products with high variance customer reviews than when they are choosing across products with low variance customer reviews.

In Study 3 we further assessed the mechanism underlying the effect of review variance on choice. Specifically, we found that participants choosing from products with high variance reviews found the reviews significantly more useful, trustworthy, and relied on them to a greater extent than participants choosing from products with low variance reviews ($p = 0.01$). These results suggest that consumers find reviews with lower variance less informative and helpful, and thus employ relative valuation strategies (e.g., compromise effect) that minimize risk (e.g., by choosing greater variety and deferring choice).

Overall, our empirical results suggest that customer review variance can act as an important moderator of various choice effects. Further, review variance also moderates how much consumers use and rely on product reviews (with higher variance reviews being rated as more useful and trustworthy). Additional studies are currently being designed to gain further insights regarding the process underlying the impact of reviews’ dispersion. This research thus contributes to our understanding of the psychological processes underlying the use of online reviews and suggests that attributes of the reviews themselves (e.g., variance) can affect how and to what extent reviews are used in consumer judgment and decision-making.

**“Don’t Tell Me What to Do!” Shoppers Rely Less on Consumer Reviews for Experiential than Material Purchases**

**EXTENDED ABSTRACT**

Consumer reviews are a pervasive form of word-of-mouth communication. We examine whether and why the type of purchase—experiential or material—affects shoppers’ reliance on consumer reviews. An archival field study and six experiments reveal that shoppers rely on reviews less when making experiential (vs. material) purchases because they believe their own evaluations (preferences, tastes, and quality judgments) are more unique from other consumers’ evaluations for experiential purchases.

Study 1 analyzed 6,508,574 Amazon.com consumer reviews written over 5 years and found that shoppers are less likely to assign a “helpful vote” to reviews for experiential purchases than for material purchases ($p < .001$). In Studies 2A-2B, participants contemplating a future experiential (vs. material) purchase (Study 2A; $N = 168$) or focusing on the experiential (vs. material) aspects of the same purchase (Study 2B; $N = 203$) reported they would rely less on consumer reviews (both $p's < .02$).

Study 3 tested the influence of positive and negative consumer reviews among participants ($N = 215$) deciding between either two cooking classes (experiential) or two coffee makers (material) for a prize in a real lottery. Participants were less likely to choose the positively-reviewed option in the experiential (vs. material) condition ($p = .02$), suggesting participants relied less on reviews when deciding on the experience.

In Study 4, participants relied less on consumer reviews for a Broadway ticket (experiential) than for speakers (material; $p < .05$), but there was no difference for company-provided information ($p = .16$), indicating the effect is specific to consumer-generated information.

In Study 5, participants ($N = 404$) asked to choose between reading consumer reviews or company-provided information were less likely to choose consumer reviews for a concert (experiential) than for a digital camera (material; $p < .001$). This was mediated by shoppers’ beliefs that their evaluations would be more unique from other consumers’ evaluations for the experiential purchase.

In Study 6, participants ($N = 469$) were presented with chips (experiential) or a flashlight (material), and they either tried or saw a picture of the product. All participants then rated the actual or perceived similarity of a randomly-selected review. Reliance on reviews was again lower for experiential (vs. material) purchases and mediated by perceived preference uniqueness; further, people overestimated how unique their evaluations were for experiential purchases.

Furthermore, in Studies 1, 5, and 6, our additional analyses suggest that people’s lesser reliance on consumer reviews for experiential purchases cannot be simply explained by the belief that evaluations of experiences are more diverse across all consumers.

In sum, shoppers are skeptical about the similarity of others’ evaluations of experiential purchases to their own, and thus rely less on consumer reviews for experiential (vs. material) purchases. This highlights that these purchases differ not only in their post-consumption enjoyment, but also in their pre-decision processes. Our findings also inform the field’s understanding of the types of purchases influenced by word-of-mouth and contribute to an emerging literature that examines the psychological processes underlying the perceived usefulness of consumer reviews. We identify the role of perceived preference uniqueness—a novel mechanism that captures the extent to which people view their own purchase evaluations as unique from others’. Extending prior research showing that consumers are more receptive to the opinions of seemingly similar others based on shared demographic characteristics, our investigation suggests that people are apt to hone in on the similarity between theirs and other consumers’ preferences, and that there are particular purchase domains in which consumers are motivated to perceive their preferences as unique, thus relying less on other consumers’ reviews.

**Consumers and Managers Reject (Superior) Algorithms Because They Fail to Compare Them to the (Inferior) Alternative**

**EXTENDED ABSTRACT**

Now more than ever, consumers and managers have the opportunity to use algorithms to make important predictions and decisions under uncertainty. For example, consumers can use recommendation
systems to decide which products to buy, use matchmaking algorithms to decide whom to date, and use algorithms to forecast the future prices of airline tickets. Professionals can use algorithms to forecast demand for products, decide which job applicants to hire, and choose which investments to make. Algorithms already outperform humans in the majority of the forecasting domains that have been tested (see Ágísdóttir et al., 2006; Camerer, 1981; Dawes, Faust, & Meehl, 1989; Grove et al., 2000; Kaufmann, Reips, & Wittmann, 2013; Kaufmann & Wittmann, 2016; Kuncel, Klieger, Connelly, & Ones, 2013; Meehl, 1954), and algorithms will become more accurate and abundant as we collect more data and develop new methods of leveraging those data. However, these algorithms cannot help consumers and managers make better predictions if they are unwilling to use them.

People are often hesitant to use forecasts from algorithms, even when they are the best alternative available. Professionals often underestimate algorithms when making predictions (e.g. Fildes & Goodwin, 2007; Sanders & Manrodt, 2003; Vrieze & Grove, 2009). Similarly, laypeople often prefer using forecasts from humans to forecasts from algorithms (Arkes et al., 2010; Diab, Pui, Yankelvich, & Highhouse, 2011; Eastwood, Snook, & Luther, 2012; Önkål et al., 2009; Promberger & Baron, 2006). However, it is still unclear what decision process leads people to use (inferior) human forecasts instead of (superior) algorithmic forecasts.

I propose that people choose between forecasting methods by (1) using their status quo forecasting method by default and (2) deciding whether or not to use the alternative forecasting method by comparing its expected performance to their performance goal. This process leads people to reject a superior algorithm when (1) the algorithm serves as their alternative forecasting method and (2) the algorithm performs better than their default forecasting method but fails to meet their performance goal. In other words, people often use human judgment by default and when deciding whether or not to switch to an algorithm they ask, “will the algorithm meet my performance goal?” instead of asking “will the algorithm outperform my default forecasting method?”. Given that algorithms only perform 10% - 13% better than human judgment on average in many forecasting domains (see Ágísdóttir et al., 2006; Grove et al., 2000) and that people often have unrealistically high expectations about forecasting performance (see Dietvorst, Simmons, and Massey, 2015), it is frequently the case that algorithms outperform human judgment but fail to reach people’s lofty performance goals.

I report the results of five studies that are consistent with the notion that people decide whether or not to use an algorithm by comparing its performance to their performance goal. In each study, participants decided whether to use their own judgment or an algorithm to complete an incentivized forecasting task. I manipulated participants’ performance goals by incentivizing them to reach different levels of performance between conditions and tested whether or not the manipulation affected participants’ choice of forecasting method.

In Study 1, participants in the “higher performance goals condition” could earn $0.40, $0.30, $0.20, or $0.10 for estimates that were off by 5, 15, 25, or 35 percentiles respectively, and participants in the “lower performance goals condition” could earn $0.20 or $0.10 for estimates that were off by 25 or 35 percentiles respectively. I found that participants in the lower performance goals condition were significantly more likely to choose to use the algorithm (69%) than those in the higher performance goals condition (52%), \( \chi^2(1, N = 544) = 15.87, p < .001 \). This finding is consistent with the notion that the majority of participants used human judgment by default and compared the algorithm’s expected performance to their performance goal when deciding whether or not to use it. In study 2, I replicated this finding, \( \chi^2(1, N = 553) = 5.14, p = .023 \), and ruled out the alternative explanation that participants in the lower performance goals condition believed that the algorithm performed significantly better than themselves.

In Study 3, participants received a summary of their own forecasting performance after making a set of 10 practice forecasts. Next, participants were assigned to one of five conditions in which they would need to achieve an absolute error of 12, 14, 16, 18, or 20 to earn a $0.25 bonus. Finally, participants chose between using the algorithm’s forecast and their own by indicating how good the algorithm’s past performance would have to be in order for them to use its forecast instead of their own. I found that participants required the algorithm’s past performance to be significantly better when they were assigned to harder incentives, \( t(506) = 2.21, p = .027 \), even though participants in each condition believed that the algorithm was the better performing alternative, \( t(506) = -3.61, p < .001 \), and participants’ estimates of the algorithm’s performance advantage was not related to their assigned condition, \( t(505) = -0.60, p = .552 \). In Study 4, I replicated the main finding from Study 3, \( t(508) = 4.01, p < .001 \), found that 66% of participants reported using their assigned incentives to decide on their requirement for the algorithm’s performance, and that the same 66% participants did not use their own past forecasting performance (i.e., the performance of the alternative) when deciding on their requirement for the algorithm’s performance.

In Study 5, I changed the participants’ default – participants used the algorithm’s forecast by default and reported how well they would need to perform in a set of 10 practice forecasts in order for them to use their own forecast instead of the algorithm’s. Under these conditions participants’ behavior reversed. Participants required their own past forecasts to be significantly better when they were assigned to harder incentives, \( z(N = 508) = 4.44, p < .001 \), and the majority of participants (64%) reported using their assigned incentives to decide on their requirement for their own performance.

REFERENCES


