Extremity Bias in Online Reviews: A Field Experiment

Leif Brandes, David Godes, and Dina Mayzlin

March 1, 2019

Abstract

In a range of studies across many platforms, submitted online ratings have been shown to be characterized by a distribution with disproportionately-heavy tails. These have been referred to as "j-shaped" or "extreme" distributions. Our focus in this paper is on understanding the underlying process that yields such a distribution. We develop a simple analytical model to capture the most-common explanations: differences in utility associated with posting extreme vs moderate reviews, and differences in base rates associated with extreme vs moderate reviews. We compare the predictions of these explanations with those of an alternative memory-based explanation based on customers forgetting about writing a review over time. The forgetting rate, by assumption, is higher for moderate reviews. The three models yield stark differences in the predicted dynamics of extremity bias. To test our predictions, we use data from a large-scale field experiment with an online travel platform. In this experiment, we varied the time at which the firm sent out a review solicitation email. Specifically, the time of review solicitation ranged between one and nine days after the end of one's vacation. This manipulation allows us to observe the extremity dynamics over an extended period both before and after the firm's solicitation email. Our results clearly support the predictions from the memory-based explanation, but are inconsistent with those from the utility-based and base-rate explanations.

1 Introduction

Word of mouth plays an important role in driving consumer decisions. In particular, existing research has shown that online reviews have a significant causal impact on purchases (Chevalier and Mayzlin (2006), Chintagunta et al. (2010)). In addition, there is a growing literature that examines the antecedents, content and consequences of online reviews.[1] One interesting robust empirical finding in the existing literature is the disproportionate prevalence of "extreme" online reviews. That is, relative to moderate review scores, extremely negative and positive review scores are posted more often, with the greatest tendency to post very high review scores. Since the highest possible rating score is often the mode, the resulting distribution has the shape of the letter 'J'. As documented by Schoenmüller et al. (2018), this 'J-shape' or extreme distribution is pervasive in a variety of categories and platforms.[2] However, the underlying behavioral mechanism driving these J-shaped distributions is not well-understood.

Why do we so often observe extreme online review distributions? One possible explanation is that the underlying distribution of experiences is relatively extreme.[3] We call this the *base*

[1] For example, see Berger (2014), Babic Rosario et al. (2016)

[2] Moe et al. (2017) recently called extreme onlne review distributions 'one of the most robust findings in product reviews' (p. 484).

[3] Hu et al. (2009) point out that there may be selection at the product purchase stage since consumers who choose to purchase the product have higher expected utility than non-purchasers. Note that this explains the higher incidence of very positive reviews and not the higher incidence of extreme negative reviews.

rate explanation since it suggests that the cause of extreme online reviews distributions is the relative high base rate of extreme experiences. Instead, it could also be the case that there is selection at the review provision stage. The most frequently-used existing explanation for extreme distributions is that consumers receive greater utility from sharing extreme opinions. For example, Anderson (1998) proposes a model of word of mouth as a function of satisfaction, where more extreme experiences increase the utility of engaging in word of mouth. We call this the *utility-based* explanation since it relies on differences in utility from providing reviews of extreme versus moderate experiences.

We propose a novel explanation behind the prevalence of extreme distributions of online reviews and test the implications of the explanation in a field test. We posit that relative to moderate experiences, extreme experiences are more emotional and arousing, which makes them more memorable. This insight is based on behavioral research that demonstrates a positive link between emotions, arousal, and memory. Specifically, researchers have found that emotional arousal improves memory encoding, consolidation and retrieval (see Kensinger (2009) for a review). However, if extreme experiences are inherently more memorable than moderate experiences, then customers have relatively more time to talk and write about them before they are forgotten. In addition, this implies that moderate experiences are forgotten more easily. Hence, this implies that extreme experiences are shared more often relative to moderate experiences. We call this the *memory-based* explanation.

Understanding the mechanism behind extreme online review distributions is important for theoretical and managerial reasons. First, it helps us understand the extent to which the distribution of online reviews is representative of the underlying distribution of consumer experiences. That is, while the base rate explanation implies that reviews are an accurate reflection of underlying experiences, the utility-based and the memory-based explanations assume that there is selection at the review provision stage. Moreover, the memory-based and the utility-based theories have very different implications on the best way to de-bias the reviews in order to obtain an accurate sense of the distribution of consumer purchase experiences. For example, if the utility-based explanation is the main driver, paying customers for reviews should decrease the bias by increasing the utility from posting moderate reviews. However, if the main driver is the memory-based explanation, a simple email reminder may be sufficient.

To empirically test the different theories, we proceed in three steps. First, we develop a simple analytical model that allows us to derive clear testable predictions for the different mechanisms. In our model, customers with and without extreme experiences arrive in each period and then decide whether to write a review of their experience. Our key novel assumption is that some consumers forget about their experience and leave the pool of potential reviewers. We separately vary the review probability, base rates, and forgetting rates for customers with and without extreme experiences. We show that while the memory-based and the base rate explanations imply extreme review distributions uniformly across time, the utility-based explanation predicts extreme review distributions only for the early periods after consumption. That is, the theories have very different predictions on the dynamics of online reviews. Since extreme distributions are so commonly-observed, this seems to suggest that the utility-based explanation is not consistent with the observed data.

Second, we examine the effect of a review solicitation email sent to all customers who have not yet written a review for their most recent experience. We show that, under certain assumptions, the competing theories make markedly different predictions for the effect of the 'reminder' email on the distribution of posted reviews: the memory-based explanation predicts a relative *decrease* in the posting of extreme experiences after a reminder, while the base rate and the utility-based explanations predict a relative *increase*. This enables us to design an empirical study to test the relative explanatory power of the different models. Finally, we report the findings from a large-scale field experiment that we designed in cooperation with a major European online travel portal where customers can book and review hotel trips. We randomly assigned customers to four different conditions that differed in the length of the time interval between the end of the customer's vacation and the reminder email. Specifically, while some customers received the email on the first day after the end of travel, others received it on the second, fifth, and ninth day. This design allows us to compare the distribution of provided reviews following a reminder email with that of a control group that did not yet receive a reminder.

Our results show that reminders lead to a relative *decrease* in the posting of extreme experiences. The effect sizes are considerable: 10 percent fewer extreme reviews are written in the treatment conditions, in which a reminder has already been received relative to the control conditions, in which the reminder has not yet been received. Importantly, this comparison holds constant across conditions the number of days that have passed since the end of travel. Accordingly, our results cannot be explained by previous work that suggests that extreme experiences for hedonic goods, such as hotels, may become more moderate over time (e.g., Moore (2012)). We then report the results from a direct estimation approach for the parameters in our model, and show that these are consistent with predictions from the memory-based explanation. Further analyses demonstrate the robustness of our results to the use of two alternative measures of review extremity. Overall, our empirical results provide strong support for our memory-based explanation and are inconsistent with the predictions from the utility-based and base rate explanations.

This paper makes several substantial contributions to the literature on online word of mouth. First, we advance existing knowledge about online product reviews by identifying the behavioral mechanism that drives the commonly-observed extreme distribution of online reviews. Specifically, ours is the first study to present a theoretical argument and empirical evidence for the memory-based explanation. The results strongly support our explanation. Our study also presents novel evidence that the distribution of reviews is not stable over time. Specifically, we show that reminding customers to write a review can change the relative extremity of posted reviews. This knowledge is highly relevant for marketers who may wish to de-bias extreme distributions. Finally, we shed first light on the highly relevant managerial question on the *timing* of solicitation emails, and show that delaying the reminder is not optimal. In contrast, previous studies focus on the content of solicitation emails, in particular, on different ways to incentivize customer review provision (e.g., Fradkin et al. (2018)).

The rest of this paper is structured as follows. In Section 2, we review previous evidence on extreme distributions, discuss existing explanations for this phenomenon, and develop the memory-based explanation from psychological and neuroscientific research on emotions and memory. In Section 3, we present our analytical model and derive testable predictions from the alternative, theoretical explanations for extreme distributions. In Section, 4, we describe our experimental design and our identification strategy. In Section 5, we present our experimental results. We discuss the implications of our findings in Section 6 and conclude in Section 7.

2 Related Literature

In this Section, we first review existing evidence on the prevalence of extreme distributions in online reviews and discuss existing explanations for this phenomenon. We then introduce a novel, memory-based explanation for extreme distributions and review psychological and neuroscientific research that demonstrates a positive link between extreme experiences, emotional arousal, and memory.

2.1 Evidence and Explanations for Extreme Distributions

Numerous studies have documented that online reviews distributions exhibit a disproportionate share of extreme reviews. This phenomenon appears in virtually all product and service categories, including books (Chevalier and Mayzlin (2006), Godes and Silva (2012), Hu et al. (2009)), DVDs (Hu et al. (2009)), movies (Dellarocas and Narayan (2006), Liu (2006)), bath, fragrance and home products (Moe and Schweidel (2012)), home improvement products (Lafky (2014)), physicians (Gao et al. (2015)), restaurants (Yelp (2018)), and accommodations (Fradkin et al. (2018)).[4]

Table 1 presents an illustrative overview of studies that found extreme distributions in online reviews and reveals three important insights. First, extreme reviews account for about two thirds

[4] Recent evidence further demonstrates that this phenomenon is not restricted to consumption experiences. For example, Marinescu et al. (2018) study online reviews on Glasdoor.com for employers, and also find a disproportionate share of extreme reviews.

of posted reviews on platforms that do not allow for reciprocal rating between buyers and sellers. Second, the highest possible rating score accounts for about fifty to sixty percent of reviews on these platforms. In contrast, platforms that allow for reciprocal ratings, such as Airbnb, exhibit an even greater share of extreme reviews, and this share is exclusively driven by extremely positive reviews (Fradkin et al. (2018)). Third, the Table reveals that most existing research has focused on Amazon reviews. In response to this over-representation of Amazon, Schoenmüller et al. (2018) recently conducted an extensive study of extreme reviews across a wide range of platforms and product categories. They report that on all 12 studied platforms that use a five-point rating scale, such as Amazon, extreme distributions are very prevalent. For Amazon itself, the authors find that between 84% to 98% of products from 24 product categories exhibit extreme distributions. In contrast, the prevalence of extreme distributions is considerably smaller for platforms that deviate from the frequently encountered five-point scale, and allow customers many more rating score options, such as RateBeer which uses a 20 point scale, or MovieLens, which uses a 10 point scale.

The prevalence of extreme distributions has led researchers to speculate on the underlying mechanism that drives this stylized fact of online reviews. The last column in Table 1 demonstrates that the utility-based explanation has by far been the most widely applied explanation for extreme distributions. In fact, we did not come across a study that a) provided an explanation for such distributions, and b) did not at least mention the idea that customers derive greater utility from sharing extreme experiences. The second most frequent explanation was the base-rate explanation, although there were relatively few mentions of this explanation. A third class of explanations related to platform-specific mechanisms, such as reciprocal-rating procedures between buyers and sellers on Airbnb (Fradkin et al. (2018)). Finally, Schoenmüller et al. (2018) discuss evidence by Mayzlin et al. (2014) and Luca (2012) on review fraud, which suggests that promotional reviews tend to be more extreme.

While empirical evidence on the relative importance of different explanations remains scarce, existing studies clearly favour the utility-based explanation for extreme distributions. For example, Schoenmüller et al. (2018) present empirical evidence from surveys, experiments, and secondary data, which shows that the utility-based explanation is the key driving factor behind extreme distributions. For example, they find that greater reviewing frequency of a customer, an inverse measure of self-selection, is associated with the posting of more moderate reviews. Similarly, forcing experimental participants to review their last product experience led to less extreme review distributions than allowing them to choose any past experience for review. While the authors also report support for other drivers, such as the base rate explanation, or review fraud, these effects are found to be much smaller. Similarly, Fradkin et al. (2018) demonstrates that the utility-based explanation is the greatest source of review bias on Airbnb. Based on a field experiment, they also

Table 1: Previous Studies that report Extreme Distributions

Study	Product Category (Data Source)	% Extreme Ratings	% Highest Rating Score	Theoretical Explanations
Chevalier and Mayzlin (2006)	Books (Amazon)	60 - 70	57 - 67	none provided
Dellarocas and Narayan (2006)	Movies (Yahoo! Movies)	65	47	utility-based
Fradkin et al (2018)*	Accommodation (Airbnb)	75	74	utility-based and others
Gao et al (2015)*	Physicians (RateMDs.com)	64	59	utility-based and others
Godes and Silva (2012)	Books (Amazon)	64	56	none provided
Hu et al. (2009)*	Books, DVDs, Video (Amazon)	58 - 64	47 - 56	utility-based, base rate
Lafky (2014)**	Home Improvement (Amazon)	-	-	utility-based, base rate
Moe and Schweidel (2012)**	Bath, Fragrance and Home (anonymous retailer)	-	-	utility-based
Schoenmueller et al (2018)*	Multiple Products and Platforms	41 - 85	31 - 84	utility-based and others
Yelp (2018)	Restaurants (Yelp)	64	48	none provided
This study	Accommodation (anonymous platform)	55	45	memory-based

Notes: To ease the comparison across studies, we re-labelled theoretical explanations as utility-bases, if authors cited Anderson (1998) as a key reference for drivers behind extreme distributions, or if they argued that posting extreme experiences yields greater utility to customers. An example are Gao et al. (2015) who introduce "hyperbole effects" in rating valence to explain the prevalence of more extreme reviews. However, most of their discussion is in the spirit of Anderson (1998), and emphasizes the greater utility that individuals derive from sharing extreme experiences. For papers marked with *, shares of rating scores for extreme distributions were manually calculated from Tables and Figures in the paper. Papers marked with ** presented evidence for extreme distributions, but did not report distributions of rating scores across categories.

show that reminder emails with $25 coupons in return for a review reduced extreme distributions relative to reminder emails without such coupons. As such coupons increase the utility of posting *any* travel experience, this finding is consistent with the idea that, in the absence of such incentives, posting extreme experiences yields greater utility for customers. Further supporting the importance of this utility-based explanation, but focusing on the impact of review costs, Lafky (2014) finds in a laboratory experiment that customers are more likely to share extreme reviews when reviewing is costly than when it is free. Overall, these studies indicate that online reviews exhibit considerable self-selection at the review-provision level.

In spite of this important insight, our understanding of the exact mechanism behind this self-selection remains limited. Specifically, it appears from the literature that customers are purely utility-directed in their decision whether to review a particular experience, and that monetary incentives serve as a straightforward way to increase review participation. However, as we argue in the following section, this perspective ignores other potential drivers behind extreme distributions, in particular the role of memory during review provision.

2.2 Extreme Experiences, Emotional Arousal, and Memory

Extreme experiences, i.e., those that involve very high or very low levels of customer satisfaction as in Anderson (1998), are frequently characterized by a high level of emotion and surprise. Indeed, research has demonstrated that customer satisfaction results from cognitive (expectancy-disconfirmation), affect, and attribution based processes (Oliver (1993)). However, if satisfaction involves a comparison between expected and actual product performance, as postulated by the expectancy-disconfirmation hypothesis, then it is - by definition - based on "prediction errors", or surprises.

Theories of learning have long acknowledged the role of surprises for memory encoding (Higgs et al. (2015)). Specifically, surprising events have been demonstrated to result in greater physiological arousal, which helps to focus attention and memory encoding, consolidation, and retrieval of such events (Kensinger (2009)). As the most basic function of memory is survival (Higgs et al. (2015)), the goal of this process is to reduce future prediction errors, thereby increasing the chances of survival. Memory thus operates by the principle of selective encoding, where the storage of the most relevant information and experiences enjoys priority as an energy- and resource-efficient way to learn. However, an adaptive memory system also needs to be flexible and allow for memory modification if necessary, e.g., through forgetting. Indeed, some authors have concluded that 'forgetting may be important for efficient use of memory, rather than a design fault' (Ward (2015), p. 220).

Besides on the level of surprise, the brain also relies on emotions to determine whether a specific memory is important enough to be kept. Importantly, this signaling function of emotions is consistent with an evolutionary perspective and holds for both, positive and negative emotions alike: while memories of negative emotions help to recognize and avoid unpleasant and dangerous situations, memories of positive emotions help to recognize and seek pleasant and rewarding situations (Nieuwenhuis (2017)). Consistent with this perspective, considerable evidence reviewed by Kensinger (2009) demonstrates that emotional experiences and stimuli are more memorable than neutral stimuli.

Overall, we conclude that there exists considerable evidence in psychology and neuroscience, which supports the idea that memory storage and learning favour extreme over moderate experiences. This implies that customers with extreme experiences have relatively more time and opportunities to share them with others before they forget. However, if moderate experiences are forgotten more easily, and shared relatively less often than extreme experiences, this behavioral pattern will result in an extreme distribution of posted reviews. The present work is the first to study this memory-based explanation for extreme distributions.

3 Theory

We proceed in three steps in this section. First, we provide a simple, analytical model of review provision for customers. We then show how existing explanations from the utility-based, base rate, and our novel memory-based theories can be integrated in this model, and study each theory's ability to explain extreme distributions. Finally, we introduce a reminder email from the review platform in our model, and derive, for each explanation, theoretical predictions for its effect on the relative arrival of extreme and moderate reviews.

3.1 A Simple Model of Review Provision

Consider the following simple review-posting process. In period $t = 0$, there are N customers who just returned home from their hotel destinations. We assume that N_x customers had an extreme experience, and N_m customers had a moderate experience. In the following, we use $i \in \{x, m\}$ to denote the type of a customer. At the beginning of each period t, each type of customer posts a review with probability r_i. Conditional on not having posted a review, a customer of type i forgets with probability ϕ_i about the experience, and permanently leaves the pool of potential posters.[5] Finally, let P_i^t represent the number of reviews of type i posted in period t and M_i^t be the number of possible "active" reviewers who have not yet posted a review or forgotten about the experience. At the start of period 1, we assume that all customers are potential reviewers: $M_i^1 = N_i$. After the first period, the expected number of active reviewers in the population must be

$$E\left[M_i^2\right] = N_i - N_i r - N_i(1-r)\phi_i = N_i\left(1-r\right)\left(1-\phi_i\right).$$

In general, we have

$$E\left[M_i^t\right] = N_i\left[\left(1-r_i\right)\left(1-\phi_i\right)\right]^{(t-1)}, \tag{1}$$

and the expected number of reviews of type i posted in period t is

$$E\left[P_i^t\right] = E\left[M_i^t\right] r_i = N_i r_i\left[\left(1-r_i\right)\left(1-\phi_i\right)\right]^{(t-1)} \tag{2}$$

Using the expression above, we can also calculate the total expected number of posted reviews

[5] For the purposes of our model, this is equivalent to a specification, in which the customer may not forget about the experience *per se*, but instead forgets to post a review about the experience.

of type i that will eventually have been posted in the long-run

$$\sum_{t=1}^{\infty} E\left[P_i^t\right] = \frac{N_i r_i}{1-(1-\phi_i)(1-r_i)}. \tag{3}$$

And, finally, we can calculate the expected number of reviews of type i posted in the first T periods:

$$\sum_{t=1}^{T} E\left[P_i^t\right] = \frac{N_i r_i \left[1-\left[(1-r_i)(1-\phi_i)\right]^T\right]}{1-(1-\phi_i)(1-r_i)} \tag{4}$$

3.2 Explaining Extreme Distributions in Reviews

In this Section, we turn to extreme distributions, and investigate what type of model can explain this stylized fact of online review distributions. Based on the reviewed literature in Section 2, we assume that extreme and moderate reviews follow a different posting process. Specifically, we explore three possible sources of heterogeneity across the two groups: 1) *the utility-based* model: a higher utility derived by the customer who posts an extreme review versus one who posts a moderate review ($r_x > r_m$), 2) *the memory-based* model: a lower forgetting rate by a customer with an extreme experience versus a customer with an extreme experience ($\phi_x < \phi_m$), and 3) *the base rate* model: a larger number of customers with extreme versus moderate experiences ($N_x > N_m$).

Our strategy is to focus on one explanation at a time. For example, in the utility-based model, we assume that $r_x > r_m$, but we also assume that $\phi_x = \phi_m \equiv \phi$ and $N_x = N_m \equiv N$. This allows us to investigate the extent to which any one source of heterogeneity could be driving the general pattern that we observe in the data. While the utility-based and base rate explanations have been offered in the literature before (although mostly informally), the memory-based model is a novel explanation.

Theorem 1 *The three explanations have different predictions for the relative number of extreme versus moderate reviews posted.*

1. *The utility-based explanation implies that the expected number of extreme reviews posted will be greater than the expected number of moderate reviews posted for low t, and the reverse for high t.*

2. *The memory-based explanation implies the same expected number of extreme and moderate reviews posted in $t = 1$ and strictly more extreme than moderate reviews posted for all $t > 1$.*

3. *The base rate explanation implies strictly more extreme than moderate reviews posted in expectation for all t.*

Proof.

$$E\left[P_x^t\right] = N_x r_x \left[\left(1-r_x\right)\left(1-\phi_x\right)\right]^{(t-1)} \tag{5}$$

$$E\left[P_m^t\right] = N_m r_m \left[\left(1-r_m\right)\left(1-\phi_m\right)\right]^{(t-1)} \tag{6}$$

1. For the utility-based explanation, we have

$$E\left[P_x^t\right] = N r_x \left[\left(1-r_x\right)\left(1-\phi\right)\right]^{(t-1)} \tag{7}$$

$$E\left[P_m^t\right] = N r_m \left[\left(1-r_m\right)\left(1-\phi\right)\right]^{(t-1)} \tag{8}$$

Note that $E\left[P_x^t\right] > \left[P_m^t\right]$ iff

$$r_x\left(1-r_x\right)^{(t-1)} > r_m\left(1-r_m\right)^{(t-1)} \tag{9}$$

$$\iff \tfrac{r_x}{r_m} > \left[\tfrac{1-r_m}{1-r_x}\right]^{t-1} \tag{10}$$

At $t = 1$, Equation (10) holds since $r_x > r_m$. The right hand side of Equation (10) is monotonically increasing in t if $r_x > r_m$. Hence, there exists a t' such that Equation (10) holds iff $t < t'$. Hence, the expected number of extreme reviews posted is greater than the expected number of moderate reviews posted for low t only. □

2. For the memory-based explanation, we have

$$E\left[P_x^t\right] = N r \left[\left(1-r\right)\left(1-\phi_x\right)\right]^{(t-1)} \tag{11}$$

$$E\left[P_m^t\right] = N r \left[\left(1-r\right)\left(1-\phi_m\right)\right]^{(t-1)} \tag{12}$$

Note that given $\phi_x < \phi_m$, $E\left[P_x^t\right] > \left[P_m^t\right]$ for all $t > 1$, and $E\left[P_x^t\right] = \left[P_m^t\right]$ for $t = 1$. □

3. For the base rate explanation, we have

$$E\left[P_x^t\right] = N_x r \left[\left(1-r\right)\left(1-\phi\right)\right]^{(t-1)} \tag{13}$$

$$E\left[P_m^t\right] = N_m r \left[\left(1-r\right)\left(1-\phi\right)\right]^{(t-1)} \tag{14}$$

Clearly, given $N_x > N_m$, the expected number of posted extreme reviews is strictly greater than the expected number of moderate reviews for all t. □

■

Theorem 1 establishes that - for each theory - there exists a certain range of low t, in which more extreme reviews will be posted in expectation than moderate reviews. Interestingly, the utility-based model (which Table 1 revealed to be the most popular explanation for extreme distributions in the literature) actually predicts a reversal of the effect for large t. What is the intuition behind the non-monotonicity in the utility-based explanation? Note that the expected number of reviews posted in each period is a product of two different variables: 1) the posting rate, and 2) the current pool of active reviewers. For low t, the fact that the extreme reviews have a higher posting rate drives the result that we expect more of them to be posted relative to moderate reviews. However, it is also the case that, over time, the population of active users is larger for moderate reviews (since there are relatively fewer of them leaving the pool due to posting). As t increases, this effect starts to dominate, which is why we expect more moderate reviews to be posted in expectation.

In contrast, the memory-based explanation implies that more extreme reviews will be posted for all $t > 1$. The intuition is the following: given that the theory assumes that extreme experiences are more memorable, we should see fewer customers with extreme experiences leaving the pool of active reviewers due to forgetting. This implies that we expect to have more extreme reviews to be posted in expectation. Finally, the base rate explanation assumes that extreme experiences occur more frequently, which means that the pool of active reviewers and posters will be greater for extreme experiences for all t.

Theorem 1 does not fully differentiate between different explanations for observed extreme review distributions, because each theory predicts the same outcome for low t. However, the fact that the utility-based explanation implies that we would actually expect to observe relatively more moderate reviews for large t may be suggestive that this theory by itself is not sufficient to explain the observed empirical phenomenon. Yet, Theorem 1 does not allow us to distinguish between the memory-based and base rate explanations. To do so, we examine a related phenomenon - the effect of a reminder email on review provision.

3.3 Reminder Analysis

In this Section, we examine the effect of a reminder email sent by the review platform on the posting of reviews of the two types of experiences. Suppose that the reminder arrives at the beginning of period $T + 1$. Here we model the reminder as affecting those customers who forgot about the experience (or, equivalently, those customers who forgot to post about the experience) in periods $t = 1, ..., T$. Specifically, we assume that the reminder brings back those customers who previously left the pool of potential posters due to forgetting. Hence, following a reminder, the expected number of possible active reviewers in the population are all the customers who have not yet posted a review:

$$E\left[M_i^{(T+1)}|\text{reminder}\right] = N_i - \sum_{t=1}^{T} E\left[P_i^t\right] = N_i\left[1 - r_i\frac{1-\left[(1-r_i)(1-\phi_i)\right]^T}{1-(1-\phi_i)(1-r_i)}\right] \tag{15}$$

In the following three subsections, we will compare what the different explanations would predict in terms of the relative proportion of expected extreme reviews if there is a reminder at the beginning of period $T+1$ in comparison to if there is not.

3.3.1 The Memory-Based Explanation

Recall that the number of available reviewers in period T+1 without a reminder is given by

$$E\left[M_i^{(T+1)}|\text{no reminder}\right] = N_i\left[(1-r)(1-\phi_i)\right]^{(T)} \tag{16}$$

According to the memory-based explanation, the expected difference in the number of reviews in period $T+1$ of type i with and without a reminder is given by:

$$\begin{aligned}\Delta_i^T &\equiv E\left[P_i^{T+1}|\text{reminder}\right] - E\left[P_i^{T+1}|\text{no reminder}\right] &(17)\\ &= Nr\left[1 - r\frac{1-\left[(1-\phi_i)(1-r)\right]^T}{1-\left[(1-\phi_i)(1-r)\right]}\right] - Nr\left[(1-r)(1-\phi_i)\right]^T &(18)\end{aligned}$$

At $T=1$, this becomes

$$\begin{aligned}\Delta_i^1 &= Nr\left[1-r\right] - Nr\left[(1-r)(1-\phi_i)\right]\\ &= Nr\left[1-r\right]\phi_i > 0 &(19)\end{aligned}$$

Since the reminder adds more potential posters, of course, this will increase the posted reviews. The key to notice is that, since $\phi_m > \phi_x$, $\Delta_m > \Delta_x$. Put differently, for the memory-based explanation, a reminder after period $T = 1$ will lead to, relatively, more moderate reviews as compared with a case in which there's no reminder. The intuition for this result is that more moderate reviews are brought back by the reminder.

Extending our analysis to the period $T = 2$, we have

$$\begin{aligned}
\Delta_i^2 &= Nr\left[1 - r\frac{1-[(1-\phi_i)(1-r)]^2}{1-[(1-\phi_i)(1-r)]}\right] - Nr\left[(1-r)(1-\phi_i)\right]^2 \\
&= N\left[r\left[1-r\left(1+\left[(1-\phi_i)(1-r)\right]\right)\right] - r\left[(1-r)(1-\phi_i)\right]^2\right] \\
&= N\left[r(1-r)(1-r(1-\phi_i)) - r\left[(1-r)(1-\phi_i)\right]^2\right] \\
&= N\left[r(1-r)\left[(1-r(1-\phi_i)) - (1-r)(1-\phi_i)^2\right]\right] \qquad (20)
\end{aligned}$$

Differentiating Δ_i^2 with respect to ϕ_i yields

$$\frac{\partial}{\partial\phi_i}\Delta_i^2 = N\left[r(1-r)\left[r + 2(1-r)(1-\phi_i)\right]\right] > 0 \qquad (21)$$

As Δ_i^2 is increasing in ϕ_i and $\phi_m > \phi_x$, the memory-based explanation predicts that a reminder after period $T = 2$ will lead to a relatively higher increase in moderate reviews than extreme reviews following a reminder as compared with a no-reminder control.

3.3.2 The Utility-Based Explanation

Starting from our definition of Δ_i^T, it is easy to check the $T = 1$ case for the utility-based explanation by just changing subscripts:

$$\begin{aligned}
\Delta_i^1 &= Nr_i\left[1-r_i\right] - r_i\left[(1-r_i)(1-\phi)\right] \\
&= Nr_i\left[1-r_i\right]\phi \qquad (22)
\end{aligned}$$

The first important thing to notice here is that a "pure utility model" (in which there's no forgetting, i.e., $\phi = 0$) would predict no difference in extremeness due to a reminder after period 1. More central to the point, we see that for $r_i \in (0, 0.5)$[6], the utility-based explanation predicts $\Delta_x > \Delta_m$. The intuition for this result is that the review process favors the posting of extreme reviews to begin with and the reminder exacerbates this by giving extreme forgetters a "second chance" at posting, which they do with higher likelihood than the second chance moderate reviewers.

Extending our analysis to the case $T = 2$, we can again just use the result from the memory-based explanation, and change subscripts:

$$\begin{aligned}
\Delta_i^2 &= N\left[r_i(1-r_i)\left[(1-r_i(1-\phi)) - (1-r_i)(1-\phi)^2\right]\right] \\
&= N\left[r_i(1-r_i)\left[1-(1-\phi)(1-\phi+r_i\phi)\right]\right]
\end{aligned}$$

[6] Since this is the periodic rate of posting, we would expect $r << 0.5$

Differentiating with respect to r_i yields:

$$\begin{aligned}\frac{\partial}{\partial r_i}\Delta_i^2 &= N\left[(1-2r_i)\left[1-(1-\phi)(1-\phi+r_i\phi)\right]-r_i(1-r_i)\phi(1-\phi)\right] \\ &= N\left[(1-2r_i)\left(1-(1-\phi)^2\right)-r_i\phi(1-\phi)(2-3r_i)\right]\end{aligned}$$

It follows that

$$\frac{\partial}{\partial r_i}\Delta_i^2 < 0 \Longleftrightarrow \frac{(1-2r_i)}{r_i(2-3r_i)} < \frac{\phi(1-\phi)}{1-(1-\phi)^2}$$

where we assume that $r_i < \frac{2}{3}$. It is easy to see that $\frac{\phi(1-\phi)}{1-(1-\phi)^2} < 1$. Moreover, $\frac{(1-2r_i)}{r_i(2-3r_i)} > 1 \Longleftrightarrow r_i < \frac{1}{3}$. This gives us a condition, under which the utility-based explanation would predict a relative increase in the posting of extreme reviews after a reminder in $T = 2$. Note that $r_i < \frac{1}{3}$ also implies $\Delta_x > \Delta_m$ for $T = 1$.

3.3.3 Base Rate Explanation

Recall that, in this model, the heterogeneity in the posting of extreme and moderate reviews is driven only by differences in the proportion of extreme and moderate reviews in the population: N_x and N_m. Here, we will go through the main results as above to see what this model would predict.

Starting from the above definition of Δ_i, and changing subscripts, Δ_i^1 is given by

$$\Delta_i^1 = N_i r(1-r)\phi \tag{23}$$

So, we see that the relationship between Δ_x and Δ_m is determined by the relationship between N_x and N_m. Note that under the standard assumption in the literature that $N_x > N_m$, we obtain that a reminder after $T = 1$ results in a relative increase in extreme reviews relative to a case without reminder.

Looking at $T = 2$, we have:

$$\begin{aligned}\Delta_i^2 &= N_i\left[r(1-r)\left[1-r(1-\phi)-(1-r)(1-\phi)^2\right]\right] &(24)\\ &= N_i\left[r(1-r)\left[1-(1-\phi)(1-\phi(1-r))\right]\right] > 0 &(25)\end{aligned}$$

Again, we see that relatively more extreme reviews will be posted in response to a reminder after $T = 2$ when $N_x > N_m$ then without a reminder.

We summarize our previous analytical results on the effect of a reminder on the relative extremeness for each of the three explanations in the following theorem:

Theorem 2 *Assuming that a reminder makes aware all previous forgetters. Then a reminder at time T has the following effect on relative extremeness: (a) According the memory-based explanation, for both $T = 1$ and $T = 2$, the reminder causes less relative extremeness; (2) According to the utility-based explanation, for both $T = 1$ and $T = 2$, the reminder causes more extremeness for $r_i < \frac{1}{3}$; (3) According to the base rate explanation, a reminder after $T = 1$ or $T = 2$ will exacerbate base rate heterogeneity. That is, if $N_x > (<) N_m$, then following a reminder, we will see (in expectation) more (fewer) extreme reviews relative moderate reviews, as compared with a case in which there is no reminder.*

4 Experimental Design and Data

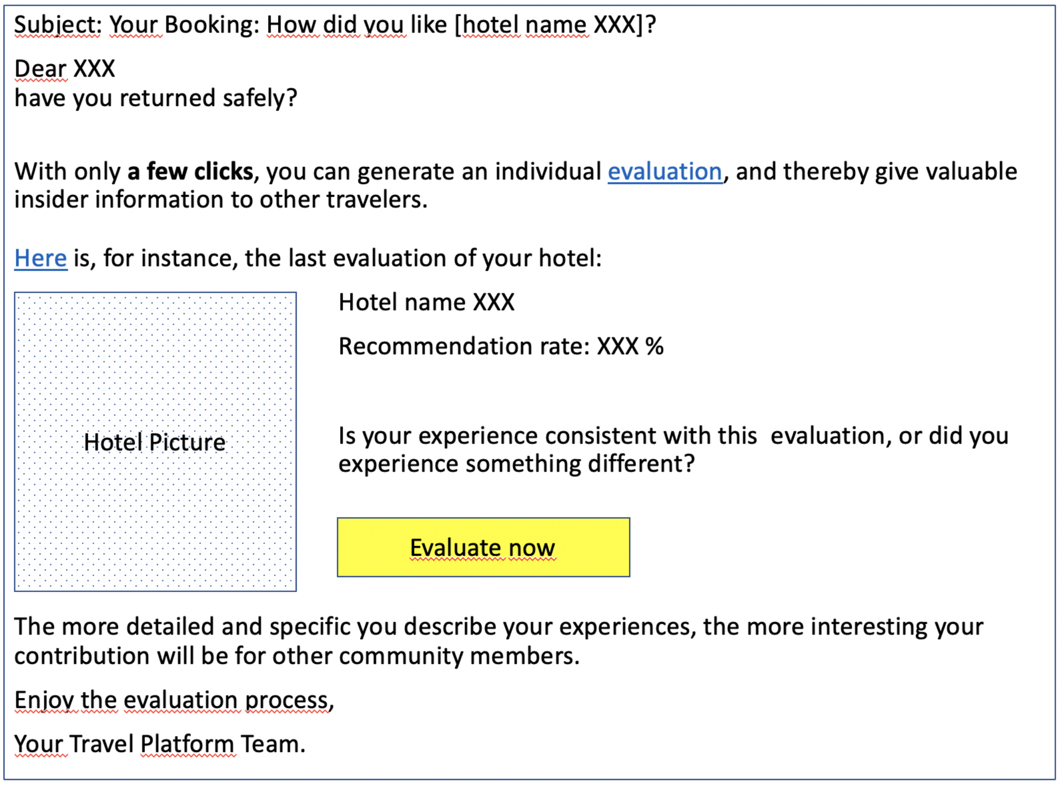
To test the theoretical predictions from the three different explanations, we conducted a field experiment in cooperation with a large, European online travel platform. The travel platform wishes to remain anonymous.

4.1 Company Background

For more than 10 years, the partnering online travel platform has been very successful, making it one of the two largest travel platforms in its core market. In the 12 months leading to the start of our experiment, the company attracted, on average, 41k hotel bookings per month, resulting in monthly revenues in excess of 61 million Euro. The platform attributes much of its success to the availability of more than 7 million customer reviews for more than 700k hotels on its site, and places great strategic importance on the currentness of its customer reviews. To reflect on the dynamic quality of hotels, the platform constructs average hotel rating scores based only on reviews from within the last two years, even if older reviews for this hotel are available.

The travel platform obtains review content from two different groups of travelers: its customers, i.e., those who have previously booked a vacation through the platform's travel agency, and travelers who have previously booked a vacation with a different travel platform, or agent. In this way, the platform combines the approaches of similar platforms, such as Expedia (where only customers of the platform can write a review), and Tripadvisor (where all travellers can submit their reviews). To receive more review content from its customers, the platform sends out a review solicitation email on the first day after the end of vacation to all customers who have not yet provided a review for their hotel experience. This email welcomes customers home, asks them for a hotel review, and provides links to the most recent evaluation for this hotel, and an online rating form. Figure 1 displays a stylized example of this email.

Figure 1: First Solicitation Email: Content and Form



Subject: Your Booking: How did you like [hotel name XXX]?

Dear XXX
have you returned safely?

With only **a few clicks**, you can generate an individual evaluation, and thereby give valuable insider information to other travelers.

Here is, for instance, the last evaluation of your hotel:

Hotel Picture

Hotel name XXX

Recommendation rate: XXX %

Is your experience consistent with this evaluation, or did you experience something different?

Evaluate now

The more detailed and specific you describe your experiences, the more interesting your contribution will be for other community members.

Enjoy the evaluation process,

Your Travel Platform Team.

Notes: A stylized example for the content and form of the first solicitation email that non-reviewers receive after the end of their vacation.

If a customer clicks on the email link to the online rating form, she will be asked to answer a number of questions, such as whether she would recommend the hotel (Yes/No), how she would rate the hotel on a scale from 1 (very bad) to 6 (very good) overall, how she would rate different quality aspects of the hotel (e.g., location, service), and how she would rate the value for money at this hotel. The consumer then needs to provide a text description that is at least 100 characters long, and is asked about some personal and travel characteristics (e.g., age, country of residence, timing and length of stay, reason for travel). If a customer does not respond to this email, the travel platforms makes up to two additional attempts to solicit a review from this customer. The second and third email (if the second email did not result in a review) are sent on the fifth and ninth day after the end of vacation, respectively.[7] If no review has been provided after 13 days, the company ends it review solicitation attempt, but sends one last email, in which customers can win a 100 Euro voucher for their next travel booking. It is important to note that the purpose of this email is only to encourage future bookings, and that the words "review" or "travel experience" are not mentioned anywhere in this email.

It was against this background that the company agreed to implement a field experiment, in

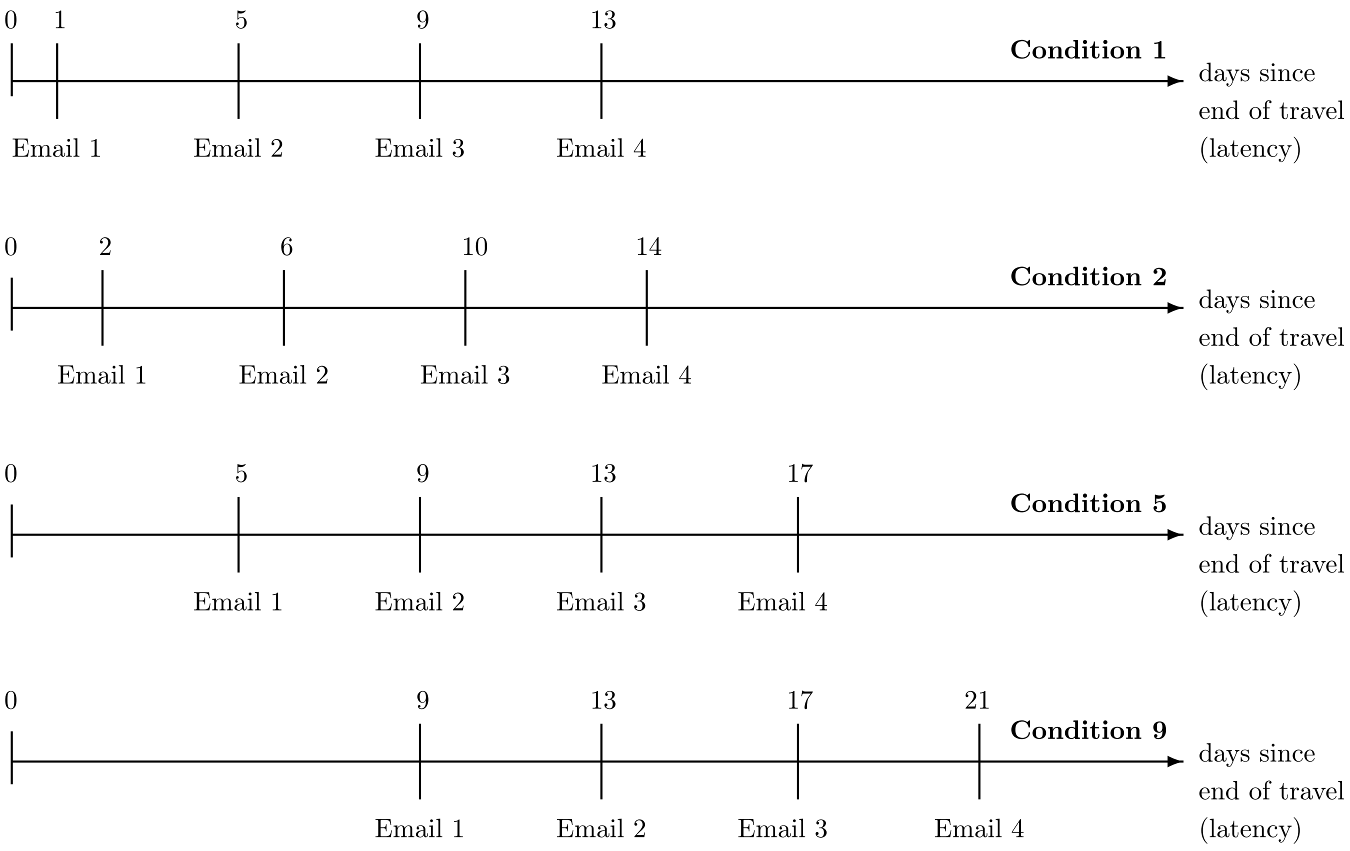
[7] The form and content of the second and third email differ from that of the first email. However, as we only focus on the effect of the first email in our empirical study, we do not discuss these differences in more detail throughout this paper.

which it randomly allocated customers to one of four experimental conditions that we designed to test for the effect of a reminder on the share of extreme reviews. Importantly, these conditions differed only in the timing of the solicitation emails. All other aspects of the emails and review solicitation procedure remained identical across conditions.

4.2 Experimental Manipulation

Figure 2 displays the full email protocol with our four experimental conditions. Condition 1 represented the previously discussed status quo at the travel platform. In the other three conditions, we increased the amount of time between the end of travel and the day that the first reminder email was sent: in condition 2, the first email was sent on the second day after the end of travel, and in conditions 5 and 9, it was sent on the fifth and ninth day after the end of travel, respectively.

Figure 2: Full Email Protocol



Notes: Displayed are the four experimental conditions that we used in our study. Condition 1 represents the current status quo.

Returning customers were randomly allocated to the four different conditions as follows. On the first day after the end of a vacation, an algorithm confirmed each customer's review status, i.e., if a review had already been provided, or not. All customers who had not yet provided a review for the vacation under study[8], were randomly allocated to one of our four conditions. Each condition had the same allocation probability. To avoid unnecessary emails to customers in conditions 2, 5, and 9, the system always confirmed a consumer's review status on the scheduled day for the first email before sending it out. The travel platform implemented this experiment on June 1, 2017 and we obtained data on sent reminder emails and review provision between June 1, 2017 to September 26, 2017. In addition, we received detailed information on bookings and hotel characteristics, which allowed us to match this information to reminder emails and hotel reviews.

Based on the previously discussed study design, there exist six different approaches to identify the effect of a reminder email on review extremity. Table 2 provides an overview of these approaches. Test 1 uses only reviews that were provided on the first day after the end of travel, and compares the share of extreme reviews across condition 1, where the reminder email has already been sent, and all other conditions, where the reminder email has not yet been sent. As the experimental treatment in our design is being reminded through notification of the reminder email, condition 1 serves as the treatment condition in Test 1, and the others serve as control conditions. In Test 2, we use more observations to increase the statistical power of our test. Specifically, we include all reviews that were posted within the first four days after the end of travel, and compare the share of extreme reviews across the treatment condition 1 and control conditions 5 and 9. Note that, since the reminder email is sent on the second day in condition 2, we can no longer use this condition in our control group.

Table 2: Experimental Design: Treatment and Control Conditions

	Latency values included in the analysis:					
	Day 1 (Test 1)	**Days 1-4 (Test 2)**	**Day 2 (Test 3)**	**Days 2-4 (Test 4)**	**Day 5 (Test 5)**	**Days 5-9 (Test 6)**
Treatment	Condition 1	Condition 1	Condition 2	Condition 2	Condition 5	Condition 5
Control	Conditions 2,5,9	Conditions 5,9	Conditions 5,9	Conditions 5,9	Condition 9	Condition 9

Table 2 shows that there exist four additional tests that cleanly assign posted reviews from a given day after end of travel to treatment and control conditions, two using condition 2 as treatment, and two using condition 5 as treatment. However, as we move from left to right in the Table, there remain fewer, non-treated observations left to serve as the control group. Note that by holding constant across conditions the number of elapsed days since the end of travel and review provision in Tests 1-6, we are able to rule out common patterns across time, such as improved customer understanding of past extreme experiences (Moore (2012)), as an alternative

[8] About five percent of reviewers in our sample provided a review before the end of their vacation, and another three percent provided a review on the day that their vacation ended.

explanation for a change in the share of extreme reviews across conditions. Ending this section, we emphasize that the focus of our study is on the effect of the *travel platform*'s reminder emails on review provision and extremity, and not on effects of the hotel management's communication with customers.

4.3 The Data

Our original data set included 200k hotel bookings that ended between June 1, 2017 and September 26, 2017. However, to have a balanced number of observations across all four experimental conditions by allowing at least nine days between the end of travel, and the last day with solicitation email information in our sample, we excluded all bookings with end dates between September 18, 2017 and September 26, 2017. Based on this approach, our final data set includes observations for 190,863 hotel bookings. Table 3 provides summary statistics on booking and customer characteristics for the complete sample. We observe that trips lasted on average about eight days, with an average price of slightly less than 1,700 Euro. Looking at customer characteristics, we see that the average trip involved 2.35 travellers (the median value was 2), that customers returned on average from one trip within our sample period (although some customers had multiple bookings), and that the average customer age was almost 41 years.[9] The final row shows that the review probability in our sample is 20 percent. Accordingly, the condition that $r_i < \frac{1}{3}$ in Theorem 2 is met, such that the utility-based explanation predicts an increase in review extremity in our data in response to a reminder email.

Table 3: Summary statistics (complete sample)

Variable	Mean	SD	Min.	Max.	N
Travel Duration	8.16	4.88	0	381	190,863
Price	1,675	1,135	0	27,758	190,863
Travellers per Booking	2.35	0.98	1	18	190,863
Bookings per Customer	1.11	0.67	1	34	190,863
Customer Age	40.94	13.22	0	88	170,566
Review Provision	0.20	0.40	0	1	190,863

To evaluate the effectiveness of our randomization procedure, we also tested for any differences across all four experimental conditions. Table 4 displays summary statistics and results from Kruskal-Wallis tests for differences across conditions. By looking at the data across conditions, we

[9] In Section 5.5, we demonstrate that our results are robust to the exclusion of customers with multiple bookings.

see very little variation, which is reassuring for the effectiveness of our randomization procedure. The results from Kruskal-Wallis tests largely support this impression, and only detect a marginally significant difference across conditions for the number of travellers per booking.

Table 4: Balance Checks Across Treatment Conditions

	Condition 1		**Condition 2**		**Condition 5**		**Condition 9**		**Kruskal-Wallis**
Variable	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Travel Duration	8.13	4.19	8.16	5.37	8.13	4.53	8.16	5.40	0.254
Price	1,673	1,140	1,676	1,145	1,668	1,126	1,674	1,134	0.660
Travellers per Booking	2.34	0.98	2.35	0.98	2.34	0.98	2.35	0.99	7.543*
Bookings per Customer	1.11	0.62	1.12	0.77	1.11	0.69	1.11	0.62	2.646
Customer Age	41.01	13.22	41.07	13.27	40.98	13.22	40.90	13.19	2.778
	$N = 46,842$		$N = 46,865$		$N = 46,699$		$N = 46,876$		

Notes: Number of observations are for Travel Duration, Price, Travellers per Booking, and Bookings per Customer. For Customer Age, the associated values are $N = 41,781$, $41,812$, $41,801$, and $41,901$. *p<0.10

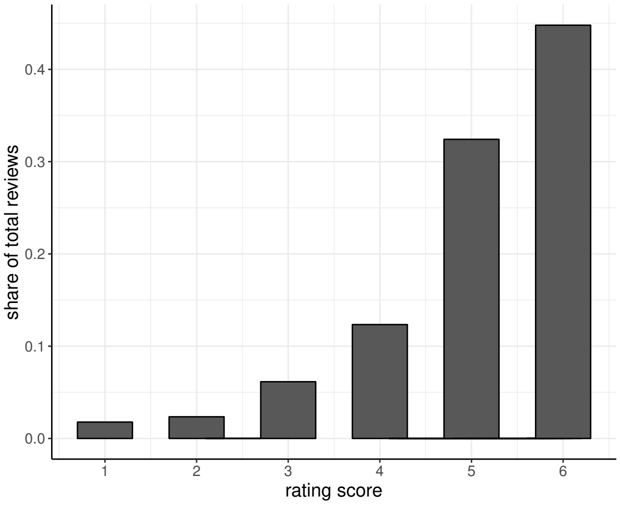
5 Empirical Results

We present our empirical results in four steps. First, we establish that our review data exhibit the well-known extreme distribution that we aim to explain. Second, we test for the underlying mechanism behind this distribution by identifying the effect of a reminder on the share of extreme ratings. Third, we present results from a direct estimation approach for our model parameters. Finally, we report the results from a number of robustness checks.

5.1 Establishing the Extreme Distribution

Figure 3 displays the rating score distribution in our sample, and yields two important insights: first, and comparable to previous research, we observe a left skewed distribution, in which 45 percent of reviews involve the highest possible rating score (6). Second, reviews with the most negative rating score (1) are extremely rare, and account for less than two percent of all posted reviews in our sample. To make the share of extreme ratings in our sample comparable to shares of around 50 to 65 percent in previous studies (as reviewed in Table 1), we classify a review to be extreme if it involves a rating score of 1,2,3, or 6. Based on this approach, extreme reviews account for 55 percent in our sample.[10]

Figure 3: Distribution of Rating Scores at Travel Platform



Notes: Displayed is the distribution of the overall rating score in posted reviews.

5.2 The Effect of a Reminder on Review Extremity

In Theorem 2 in section 3.3 we summarized that the memory-based explanation predicts a decrease in the share of extreme reviews in response to an email reminder, whereas the utility-based and the base rate explanations predict an increase in the share of extreme reviews after an email reminder.

Table 5 presents the results on the average treatment effect of a reminder email on review extremity for our six identification approaches. In Tests 1 and 2, we see that the share of extreme reviews is significantly lower in Condition 1, than in the other Conditions. Specifically, in Test 1, we focus on the first day after the end of travel and find that the share of extreme reviews is 55 percent in Condition 1 but 61 percent when pooling Conditions 2, 5 and 9. Similarly, Test 2 shows that across days 1 to 4, the share of extreme reviews is 54 percent in condition 1, but 61 percent when pooling conditions 5 and 9. Looking at Tests 3 and 4, we see that the share of extreme reviews is also significantly lower in Condition 2 on days 2 to 4 after the end of travel than in Conditions 5 and 9. The results for Tests 5 amd 6 replicate this pattern for days 5 to 8 after the end of travel when comparing Condition 5 to Condition 9. Overall, the displayed results strongly support our novel memory-based explanation, and contradict the predictions of the utility-based and base rate explanations. We summarize this finding as our first result.

Result 1 *Immediately after an email reminder, the share of extreme reviews reduces significantly. This pattern is consistent with the predictions of the memory-based explanation, but inconsistent with predictions of the utility-based and base-rate explanations.*

[10] As we report in Section 5.5, our results are robust to alternative extremeness classifications, in which we either exclude rating scores of 3, or also include rating scores of 4. These two classifications imply a share of 49 and 67 percent extreme ratings, respectively.

Table 5: Share of Extreme Reviews Across Conditions

	Condition 1			Condition 2			Condition 5		
	vs.			vs.			vs.		
	Conditions 2, 5, and 9			Conditions 5 and 9			Condition 9		
Set of Days	1	2,5 and 9	z-stat.	2	5 and 9	z-stat.	5	9	z-stat.
Test 1: Day 1	0.55	0.61	3.52***						
Test 2: Days 1 to 4	0.54	0.61†	5.26***						
Test 3: Day 2				0.56	0.62	2.79***			
Test 4: Days 2 to 4				0.55	0.60	3.77***			
Test 5: Day 5							0.56	0.63	1.65*
Test 6: Days 5 to 8							0.55	0.60	2.33**

Notes: Displayed are proportions of extreme reviews across sets of days after end of travel and conditions. z-Stat. denotes z-statistic for tests of proportion equality across conditions. † This value is based only on Conditions 5 and 9. *** $p<0.01$, ** $p<0.05$, *$p<0.10$

Table 6 displays estimation results from Logit models on the likelihood of an extreme review when controlling for the number of travellers per booking. This model specification addresses the potential concern that the previously reported differences across conditions might be a reflection of the previously detected differences in the number of travellers per booking across conditions (as displayed in Table 4). However, the results in Table 6 clearly reject this idea, and demonstrate that the likelihood of extreme reviews is consistently around 6 percent lower when travellers have just received a reminder email, than when they have not. Overall, these results confirm our previous insights and provide strong support for our memory-based explanation as an important driver behind extreme distributions.

While our theoretical model does not make any predictions about whether an email reminder affects both ends of the review scale identically or differently, we acknowledge that this question is of great managerial relevance. Figure 4 therefore illustrates the effect of reminders on the distribution of review scores in more detail. Specifically, the figure shows the distributions of rating scores across treatment and control conditions for all six comparisons from before. In panel a), for example, the focus is on the first day after the end of travel. Accordingly, the treatment condition consists of Condition 1, and the control conditions are Conditions 2, 5, and 9. In line with our previous results, we see that extreme reviews account for a lower share of reviews in the treatment condition. However, Figure 4 provides the novel insight that this effect is driven by changes on both ends of the review scale. Indeed, when looking at panels b) to f), we see that this pattern is consistent across all six comparisons. To determine the implications of this pattern for the overall rating scores, we compared overall rating scores across treatment and control conditions for our six test scenarios. Table 7 shows that, with the exception of Day 1, there is no significant

Table 6: Logit Estimations for Likelihood of Extreme Reviews Across Conditions

	Condition 1 vs. Conditions 2, 5, and 9		Condition 2 vs. Conditions 5 and 9		Condition 5 vs. Condition 9	
Variable	Test 1: Day 1	Test 2: Days 1-4†	Test 3: Day 2	Test 4: Days 2-4	Test 5: Day 5	Test 6: Days 5-8
Condition 1 (Treatment)	-0.059*** (0.017)	-0.062*** (0.012)				
Condition 2 (Treatment)			-0.059*** (0.021)	-0.052*** (0.014)		
Condition 5 (Treatment)					-0.072* (0.042)	-0.057** (0.024)
Travellers per Booking	-0.011 (0.009)	-0.009 (0.006)	0.015 (0.009)	0.001 (0.007)	0.030*** (0.011)	0.012 (0.008)
N	3,917	7,845	3,039	6,007	2,019	4,149
Wald	14.05***	29.75***	10.39***	9.88***	14.18***	7.72**
- LL	-2,671.66	-5,354.06	-2,066.38	-4,108.27	-1,376.98	-2,848.52

Notes: Displayed are marginal effects for Logit specifications. Robust standard errors are displayed in parentheses. † This effect is measured relative to Conditions 5 and 9. *** p<0.01, ** p<0.05, *p<0.10

difference in the average review score across conditions. These results demonstrate that reminder emails can help firms to improve overall rating scores, but that this effect is rather short-lived.

We emphasize that, besides having managerial relevance, these findings also provide evidence against an alternative explanation for our reported treatment effect, namely that this effect is driven by a systematic change in overall review scores in response to the reminder. Our data, however, do not support this idea. Specifically, Tests 2-6 detected significant reductions in review extremity with the treatment, *although* there was no significant change in the overall rating score. Therefore, the reminder email does not seem to uniformly shift customers' evaluation scores, which limits the explanatory power of this alternative explanation for our findings.

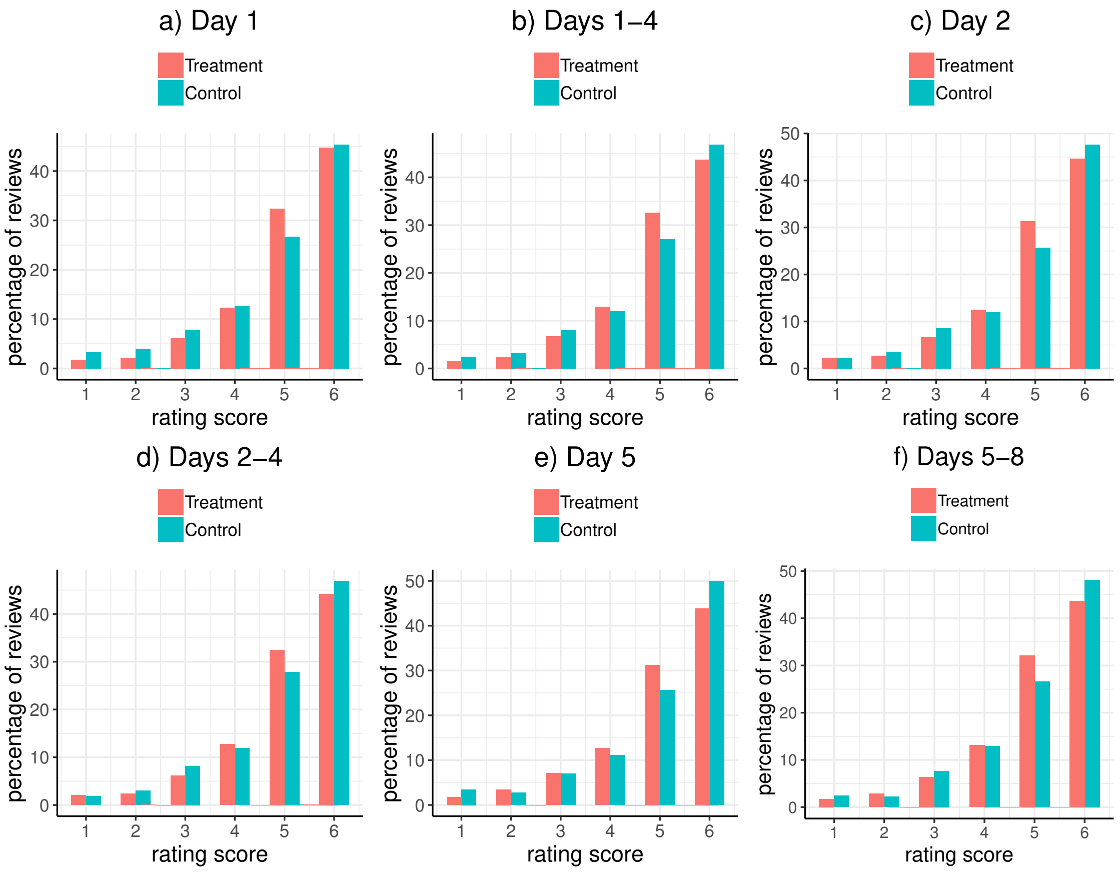
5.3 Direct Estimation Procedure

The previous results provide compelling evidence for our memory-based explanation. However, based on our simple model of review provision, we are also able to estimate the model parameters directly, as we now show.

Recall that the expected number of reviews of type i posted in period t is given by

$$E[P_i^t] = N_i r_i[(1 - r_i)(1 - \phi_i)]$$

Figure 4: Distributions of Review Scores Across Set of Days and Conditions



Notes: Displayed is the rating distribution for different sets of days after consumption. Treatment indicates conditions, in which travellers had received an email reminder prior to, or on the chosen set of days. Control indicates conditions, in which travellers had not yet received an email reminder.

By taking the logarithm on both sides, we obtain

$$\begin{aligned} & \log(E[P_i^t]) = \log(N_i) + \log(r_i) + (t-1) \cdot \log([(1-r_i) \cdot (1-\phi_i)]) \\ \iff \quad & \log(E[P_i^t]) = \nu_i + t \cdot \underbrace{\log([(1-r_i) \cdot (1-\phi_i)])}_{\beta_i}, \end{aligned} \tag{26}$$

where $\nu_i \equiv \log(N_i) + \log(r_i) - \log([(1-r_i)(1-\phi_i)])$. Equation (26) resembles a linear regression function with coefficient β_i and intercept ν_i. Importantly, the three different explanations make different predictions for the relative size of β_m and β_x. As β_i does not depend on N_i, the base rate explanation predicts that $\beta_x = \beta_m$. In contrast, the utility-based explanation predicts that $\beta_x < \beta_m$, because $(1-r_x) < (1-r_m)$. Finally, the memory-based explanation predicts that $\beta_x > \beta_m$, because $(1-\phi_x) > (1-\phi_m)$.

To estimate the coefficients β_x and β_m jointly from the same model, we proceed as follows. First, for each review type $i, i \in \{m, x\}$, we count the number of posted reviews for each latency value t, and for each experimental condition. By separating observations across conditions, we are able to control for condition-specific effects in our analysis below. Second, we replace the left-hand

Table 7: Average Review Score Across Conditions

	Condition 1 vs. Conditions 2, 5, and 9			Condition 2 vs. Conditions 5 and 9			Condition 5 vs. Condition 9		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Set of Days	1	2,5 and 9	t-stat.	2	5 and 9	t-stat.	5	9	t-stat.
Test 1: Day 1	5.04	4.92	-2.81***						
Test 2: Days 1 to 4	5.04	$4.99^{\dagger}$	-1.60						
Test 3: Day 2				5.01	4.99	-0.49			
Test 4: Days 2 to 4				5.02	5.04	-0.43			
Test 5: Day 5							5.03	4.99	0.29
Test 6: Days 5 to 8							5.04	5.02	0.19

Notes: Displayed are average review scores across sets of days after end of travel and conditions. t-Stat. denotes t-statistic for tests of mean equality across conditions. † This value is based only on Conditions 5 and 9. *** p<0.01, ** p<0.05, *p<0.10

side of equation (26) with the observed, logarithmic number of reviews, $\log(P_i^t)$. This yields the following regression equation:

$$\log(P_i^t) = \sum_{i \in \{m,x\}} \nu_{cond,i} + \beta_{extreme} \cdot \mathbb{1}_{\{i=x\}} \cdot t + \beta_{moderate} \cdot \mathbb{1}_{\{i=m\}} \cdot t + \epsilon_i^t, \tag{27}$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function that takes on the value 1 if condition $(\cdot)$ is met, and 0 otherwise.

Before estimating this equation, a short discussion of the set of days that we include in the estimation is warranted. While it would be preferable to include only observations right after the first email from each condition as in the previous subsection, such an approach is not feasible for the estimation of equation (27). The reader will recall from Section 4 that the company makes up to three attempts to solicit a review from travellers, and that the second email arrives already four days after the first email. As equation (27) specifies the relationship between the number of posted reviews of each type and time on a daily level, an exclusive use of the few days between the first and second email would thus result in a considerable lack of power to test for the equality of β_x and β_m. To use a relatively longer time-span, during which there is minimal interference in the review provision process, we thus include only observations with latency values after the third solicitation email in each condition, and reset latency values, such that a value of 1 corresponds to the first day after that email. This approach is justified by our knowledge that the travel platform's fourth email is the only email after the end of vacation that does not contain a request or incentive for review provision.[11]

[11] Further support for our approach comes from Figure A.6 in the Appendix, which shows that the fourth email is the only email that has no detectable effect on the number of posted reviews in each condition.

Table 8 displays estimation results from three different model specifications for equation (27). In each model, we include random effects to control for condition-specific differences, such as differences in the starting time of the observation period. From Model 1 to Model 3, we subsequently increase the maximum number of days after the last solicitation email from 20 to 30, where the number of observations is given by *number of days* × *4 conditions* × *2 review types*.[12] In line with the predictions from our memory-based explanation, all three models reveal that β_x is estimated to be larger than β_m. Subsequent Wald tests of coefficient equality support this observation, and reject the null hypothesis of coefficient equality in each model. Overall, the results of this direct estimation procedure provide further support for the role of our memory-based explanation for extreme distributions.

Table 8: Estimation Results on β_x and β_m

	Model 1	**Model 2**	**Model 3**
Variable	Latency up to 20 Days	Latency up to 25 Days	Latency up to 30 Days
β_x	-0.150***	-0.142***	-0.133***
	(0.008)	(0.006)	(0.005)
β_m	-0.185***	-0.159***	-0.145***
	(0.008)	(0.006)	(0.005)
N	160	198	236
Wald	766.64***	986.40***	1267.67***
Hausman Test (= FE)	0.20	0.25	2.03
Test of Coefficient Equality	13.29***	5.11**	6.92***
Overall R^2	0.81	0.82	0.84

Notes: Displayed are estimation results from a random effects model. Robust standard errors are displayed in parentheses. *** p<0.01, ** p<0.05, *p<0.10

5.4 Inferring the probabilities to post from the data

Note: I think this would be a good place for Dave's calculations on base rates, because we would use results from the previous Table for illustration.

But we need to discuss this first as I wrote in my email from Jan 15, 2019.

5.5 Robustness Checks

In this section, we report the robustness of our main results across two alternative classifications of review extremity, as well as when focusing only on those travellers who appear only once in our sample.

[12] In Models 2 and 3, the actual number of observations is slightly lower, because of a few days without posted reviews of a particular type, and the fact that we use the logarithmic number of posted reviews as dependent variable.

5.5.1 Alternative Measures of Review Extremity

In a first alternative definition, we use a more restrictive extremity definition, and exclude rating scores of 3 from the extreme definition. Overall, this approach results in the classification of 49 percent of reviews in our sample as extreme reviews. Table 9 shows the results from a replication of our main analysis, and reveals that our previous findings are robust to the use of this more restrictive measure, although the level of statistical significance is slightly reduced.

Table 9: Robustness Check: Share of Extreme Reviews Across Conditions (Restrictive Extremity Measure)

	Condition 1 vs. Conditions 2, 5, and 9			Condition 2 vs. Conditions 5 and 9			Condition 5 vs. Condition 9		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Set of Days	1	2,5 and 9	z-stat.	2	5 and 9	z-stat.	5	9	z-stat.
Test 1: Day 1	0.48	0.53	2.83***						
Test 2: Days 1 to 4	0.48	$0.53^{\dagger}$	4.20***						
Test 3: Day 2				0.49	0.53	2.79*			
Test 4: Days 2 to 4				0.49	0.52	2.32**			
Test 5: Day 5							0.49	0.56	1.67*
Test 6: Days 5 to 8							0.48	0.53	1.79*

Notes: Displayed are proportions of extreme reviews across sets of days after end of travel and conditions. In these robustness checks, a review was classified as 'extreme' if it involved a rating score of 1,2 or 6. z-Stat. denotes z-statistic for tests of proportion equality across conditions. † This value is based only on Conditions 5 and 9. *** p<0.01, ** p<0.05, *p<0.10

In a second alternative definition, we use a more comprehensive extremity definition, and also include rating scores of 4 in the extreme definition. Overall, this approach results in the classification of 67 percent of reviews in our sample as extreme reviews. Table 10 presents the results from a replication of our main analyses, and shows that the share of extreme reviews is still significantly higher in treatment versus control conditions for 5 out of 6 comparisons.

5.5.2 Results for One-Time Customers

Table 3 revealed that some customers had more than one trip that ended during our experimental period. As some of these travellers may have been allocated to more than just one experimental condition, we also conducted our main analyses when excluding customers with multiple bookings from our analysis. Table 11 presents the associated estimation results, and shows that our main results are robust to the use of this alternative sample.

Table 10: Robustness Check: Share of Extreme Reviews Across Conditions (Extensive Extremity Measure)

	Condition 1 vs. Conditions 2, 5, and 9			Condition 2 vs. Conditions 5 and 9			Condition 5 vs. Condition 9		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Set of Days	1	2,5 and 9	z-stat.	2	5 and 9	z-stat.	5	9	z-stat.
Test 1: Day 1	0.68	0.73	3.54***						
Test 2: Days 1 to 4	0.57	0.73†	4.86***						
Test 3: Day 2				0.69	0.74	2.75***			
Test 4: Days 2 to 4				0.68	0.72	3.78***			
Test 5: Day 5							0.69	0.74	1.38
Test 6: Days 5 to 8							0.68	0.73	2.42**

Notes: Displayed are proportions of extreme reviews across sets of days after end of travel and conditions. In these robustness checks, a review was classified as 'extreme' if it involved a rating score of 1,2,3,4 or 6. z-Stat. denotes z-statistic for tests of proportion equality across conditions. † This value is based only on Conditions 5 and 9. *** p<0.01, ** p<0.05, *p<0.10

6 Discussion

In this paper, we introduce a novel, memory-based, behavioral mechanism that explains the prevalence of extreme distributions, 'one of the most robust findings in product reviews' (Moe et al. (2017). p. 484). Starting from a simple model of review provision, we showed analytically that, under certain conditions, this memory-based explanation gives rise to markedly different review patterns after an email reminder than the utility-based and base-rate explanations that existing studies have focused on: while the latter both predicted a relative increase in review extremity in response to a reminder, the memory-based explanation predicted a decrease. The results from a large-scale, field experiment that we conducted in cooperation with a leading, European travel platform, showed that email reminders *decrease* the share of extreme reviews (by about 10 percent in our main specifications). Importantly, this result was robust to the use of two alternative measures of review extremity, and different estimation samples. Overall, our study thus provides first evidence for the importance of memory-based effects for review provision in the field.

6.1 Theoretical Contributions

Our work contributes to the literature on word of mouth in three important ways. First, we introduce a novel, memory-based explanation for extreme distributions, and show that this mechanism explains the observed empirical patterns better than existing explanations that focus exclusively on reviewer utility from posting, or differences in customer base rates across types of experiences. This novel mechanism was grounded in research on the positive influence of surprise and emo-

Table 11: Robustness Check: Share of Extreme Reviews Across Conditions (One-Time Customers)

	Condition 1 vs. Conditions 2, 5, and 9			Condition 2 vs. Conditions 5 and 9			Condition 5 vs. Condition 9		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Set of Days	1	2,5 and 9	z-stat.	2	5 and 9	z-stat.	5	9	z-stat.
Test 1: Day 1	0.55	0.61	3.82***						
Test 2: Days 1 to 4	0.55	0.61†	5.32***						
Test 3: Day 2				0.56	0.62	2.83***			
Test 4: Days 2 to 4				0.55	0.60	3.54***			
Test 5: Day 5							0.56	0.64	1.77*
Test 6: Days 5 to 8							0.55	0.61	2.50**

Notes: Displayed are proportions of extreme reviews across sets of days after end of travel and conditions. In these robustness checks, only customers with a single, ending trip during our sample period were included. z-Stat. denotes z-statistic for tests of proportion equality across conditions. † This value is based only on Conditions 5 and 9. *** p<0.01, ** p<0.05, *p<0.10

tional arousal - both arguably more likely to be present for extreme experiences than for moderate experiences - on memory encoding and retrieval (Kensinger (2009)), thereby predicting a greater likelihood to forget for moderate experiences. While personal experience and previous research provide evidence that consumers regularly forget to do things that they were planning to do (e.g., buying a particular item, Fernandes et al. (2016)), ours is the first study to analyze the antecedents and implications of forgetting for the provision and content of online reviews.

Second, we identify reminder emails as a novel driver for the instability of online review distributions over time. From a theoretical point of view, this contributes to our understanding of dynamics in online reviews. Existing work has either focused on social and temporal dynamics in review distributions (Li and Hitt (2008); Wu and Huberman (2008); Moe and Trusov (2011); Moe and Schweidel (2012); Godes and Silva (2012), or on dynamics that result from (hotel) managers' responses to previous reviews (Chevalier et al. (2018), Proserpio and Zervas (2017)). From a methodological point of view, we emphasize that the documented change in the share of extreme reviews in the distribution prior and after reminder emails would not have been discernible from an exclusive focus on average rating scores, as has been common in those previous studies. Future research on review dynamics should thus consider a broader range of distributional features.

Third, we contribute to the existing knowledge of biases in online reviews. While some theories, such as the base rate explanation, imply that posted reviews are an unbiased representation of customers' underlying product experiences, others, such as our novel, memory-based explanation, suggest otherwise. Against the results of our field experiment, we conclude that online reviews do not represent an unbiased view of customers' actual experiences. Our results thus confirm the importance of reviewer self-selection as a source of review bias as previously discussed in the literature (e.g., Moe and Schweidel (2012); Godes and Silva (2012)), but specify a novel mechanism for this self-selection. At the same time, the lack of empirical support for the utility-based explanation provides the novel insight that expensive, financial incentives may be unnecessary to reduce this bias, and that a simple reminder email may be sufficient. Future research could consider alternative approaches to remind customers, and study the effectiveness of these approaches under different conditions (e.g., product categories, and types of experiences).

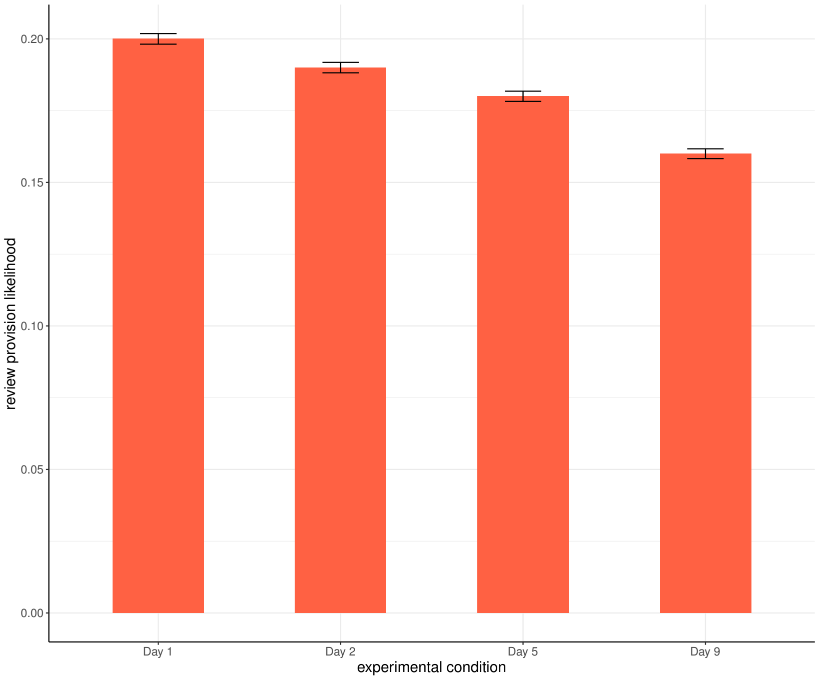
6.2 Managerial Implications

Companies are constantly looking for ways to attract more word of mouth activity from their customers. This is evident from the considerable amounts of money that they are spending on review solicitation (Babic Rosario et al. (2016)), and the increasing supply of websites that offer guidance on the best ways to get more reviews and word of mouth. Our research has three important implications for companies trying to attract more reviews from their customers. First, our identification of a memory-based mechanism behind extreme distributions implies that managers should take previous managerial recommendations that focus exclusively on interventions to change customers' cost-benefit evaluations during review provision with a pinch of salt: unless a consumer remembers to write a review in the first place, any such improvements will necessarily have to remain ineffective. Thus, firms need to establish an effective *reminder management* in order to maximize the returns from improvements in the review funnel design.

Second, our study informs managers about the optimal start point of such a reminder management. Figure 5 shows the likelihood to write a review across our four experimental conditions. We can see that waiting to remind customers is not a good strategy for travel platforms that aim to maximize the review provision likelihood from customers: while this likelihood is 20 percent when the first reminder email is sent on the first day after the end of travel, it monotonically decreases with additional delay in the first email, to only 16 percent when the first reminder email is sent on the ninth day. This shows that it is important to engage customers early on in review provision for travel experiences. Future research could consider the extent to which similar results can be observed for other product categories, in which the consumption experience has a clear start and end date that is observable to the firm (e.g., flights, cabs, restaurant visits, hospitalization, education), and how the optimal start time for reminders differs (if at all), when consumption starts and ends only after purchase (e.g., for books, DVDs, or household appliances) and cannot be observed by firms.

Finally, our study informs managers about the importance to create memorable customer experiences for review provision. In particular, our results suggest that firms that do not provide

Figure 5: Review Provision Likelihood Across Conditions



Notes: Displayed is the review provision likelihood across the four experimental conditions.

emotionally arousing, and surprising experiences will have greater difficulty to attract organic user-generated content, because fewer customers will remember these experiences over an extended time period, and thus have fewer opportunities to share them.

6.3 Limitations of this study

Just like any other research, our study is not without limitations. First, we acknowledge that our experimental design does not allow us to rule out *any* impact of the utility-based and base rate explanations on review provision. Accordingly, we emphasize that our results should *not* be read to imply that the utility-based and base rate theories are irrelevant for real-world review provision behavior. Instead, our results demonstrate that our memory-based theory possesses unique explanatory power for the existence of extreme distributions that extends beyond the explanatory power of existing theories.

Second, we acknowledge that it would have been very interesting to separate the effect of merely receiving a reminder email (but not opening it) from the effect of receiving and opening this email. Unfortunately, our data do not allow us to study this effect. The reason is that, for customers who did not open the reminder email, we have no way to establish the exact point in time when they first noted the reception of this email. This creates a substantial problem for our identification approaches that focus on review provision shortly after the email reminder was actually *sent*: for those customers that ended up writing a review on the day of the reminder email without having opened this email (around 0.6 percent of reviewers), we do not know, whether they had already noted the reception of the reminder email, and thus been reminded, or not. However, a closer look at our data suggests that, even with this information, the effect would have been very difficult to identify as this group of customers is extremely small, and accounts for only 0.6 percent of all reviewers in our sample.

7 Conclusion

Extreme distributions are a stylized fact of many online review distributions. In this paper, we introduced a novel, memory based mechanism to explain such distributions, and demonstrated its empirical relevance in the context of a large-scale, field experiment in the travel industry. Starting from a simple model of review provision, we showed how to integrate different theories for extreme distributions into the same theoretical framework, and derived testable predictions from this framework. Based on a specifically designed field experiment, we reported the results from six alternative identification approaches to test those predictions with our data. While our memory-based theory explained the observed empirical patterns very well, existing explanations alone were insufficient to do so. Future research should thus integrate considerations and implications of memory formation and retrieval for review provision.

References

Anderson, E. (1998). Customer satisfaction and word of mouth. *Journal of Service Research*, 1(1):5–17.

Babic Rosario, A., Sotgiu, F., De Valck, K., and Bijmolt, T. H. A. (2016). The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. *Journal of Marketing Research*, 53:297–318.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, 24(4):586–607.

Chevalier, J. A., Dover, Y., and Mayzlin, D. (2018). Channels of impact: User reviews when quality is dynamic and managers respond. *Marketing Science*.

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

Chintagunta, P. K., Gopinath, S., and 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.

Dellarocas, C. and Narayan, R. (2006). A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statistical Science*, 21(2):277–285.

Fernandes, D., Puntoni, S., van Osselaer, S. M. J., and Cowley, E. (2016). When and why we forget to buy. *Journal of Consumer Psychology*, 26(3):363–380.

Fradkin, A., Grewal, E., and Holtz, D. (2018). The Determinants of Online Review Informativeness: Evidence from Field Experiments on Airbnb. Working Paper.

Gao, G., Greenwood, B. N., Agarwal, R., and McCullough, J. S. (2015). Vocal minority and silent majority: How do online ratings reflect population perceptions of quality. *MIS Quarterly*, 39(3):565–589.

Godes, D. and Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473.

Higgs, S., Cooper, A., Lee, J., and Harris, M. (2015). *Biological Psychology*. Sage.

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

Kensinger, E. A. (2009). Remembering the details: Effects of emotion. *Emotion Review*, 1(2):99–113.

Lafky, J. (2014). Why do people rate? theory and evidence on online ratings. *Games and Economic Behavior*, 87:544–570.

Li, X. and Hitt, L. (2008). Self-Selection and Information Role of Online Product Reviewsl. *Information Systems Research*, 19(4):456–474.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3):74–89.

Luca, M. (2012). Reviews, reputation, and revenue: The case of yelp.com. Harvard Business School Working Paper.

Marinescu, I., Klein, N., Chamberlain, A., and Smart, M. (2018). Incentives Can Reduce Bias in Online Reviews. Working Paper.

Mayzlin, D., Chevalier, J., and Dover, Y. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–55.

Moe, W. and Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research*.

Moe, W. W., Netzer, O., and Schweidel, D. A. (2017). Social media analytics. In *Handbook of Marketing Analytics*, pages 483–504. Springer.

Moe, W. W. and Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, 31(3):372–386.

Moore, S. G. (2012). Some things are better left unsaid: How word of mouth influences the storyteller. *Journal of Consumer Research*, 38(6):1140 − 1154.

Nieuwenhuis, Ingrid, L. C. (2017). Memory. In *Consumer Neuroscience*, pages 133–150. MIT Press.

Oliver, R. L. (1993). Cognitive, affective, and attribute bases of the satisfaction response. *Journal of Consumer Research*, 20:418–430.

Proserpio, D. and Zervas, G. (2017). Online reputation management: Estimating the impact of management responses on consumer reviews. forthcoming in Marketing Science.

Schoenmüller, V., Netzer, O., and Stahl, F. (2018). The Extreme Distribution of Online Reviews: Prevalence, Drivers and Implications. Working Paper.

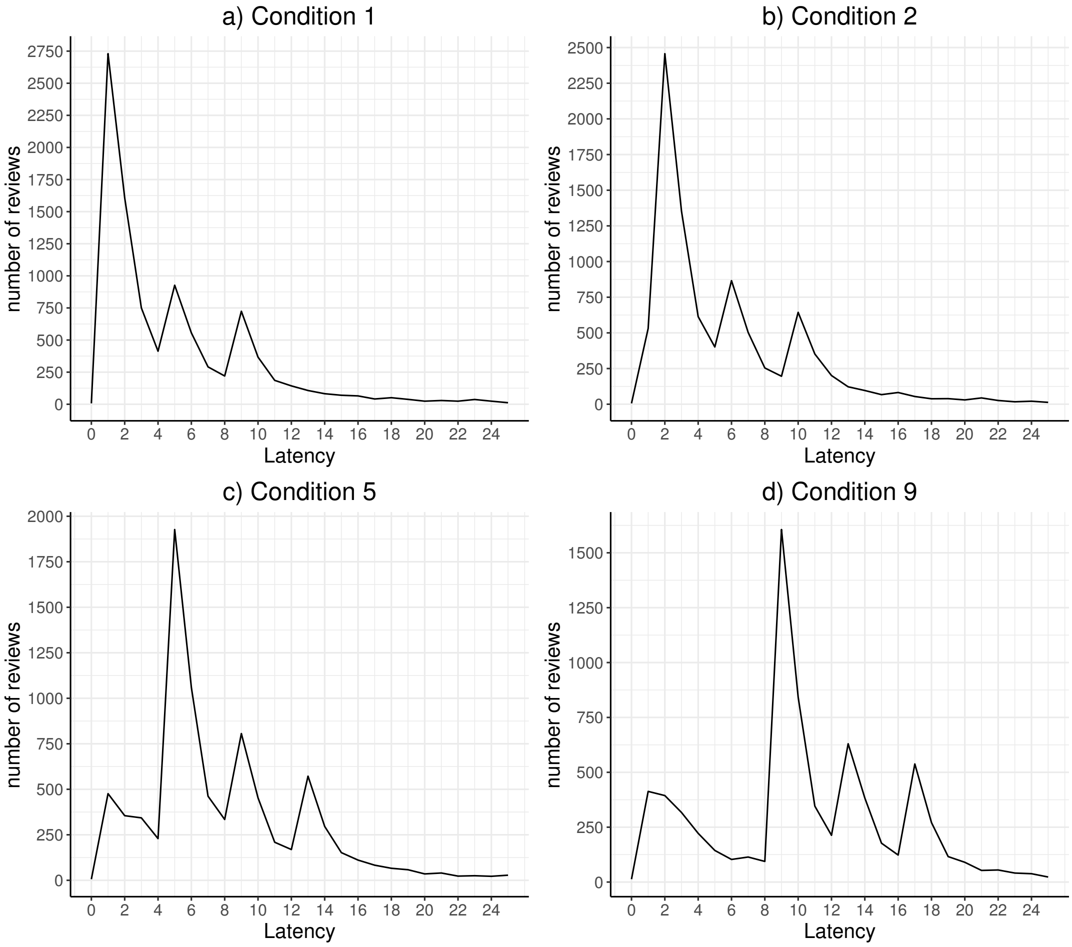
Ward, J. (2015). *The Student's Guide to Cognitive Neuroscience*. Psychology Press, third edition edition.

Wu, F. and Huberman, B. (2008). *Internet and Network Economics. Lecture Notes in Computer Science*. Springer.

Yelp (2018). Yelp factsheet 30, june 2018. https://www.yelp.com/factsheet. Accessed: 2018-08-09.

A Web-Appendix

Figure A.6: Distribution of Posted Review Volume Across Sets of Days and Conditions



Notes: Displayed is the number of posted reviews for different sets of days after consumption.