AMERICAN MARKETING ASSOCIATION

Article

Journal of Marketing Research
2020, Vol. 57(5) 853-877

DOI: 10.1177/0022243720941832
journals.sagepub.com/home/mrj
SAGE

The Polarity of Online Reviews: Prevalence, Drivers and Implications

Verena Schoenmueller, Oded Netzer, and Florian Stahl

Abstract

In this research, the authors investigate the prevalence, robustness, and possible reasons underlying the polarity of online review distributions, with the majority of the reviews at the positive end of the rating scale, a few reviews in the midrange, and some reviews at the negative end of the scale. Compiling a large data set of online reviews—over 280 million reviews from 25 major online platforms—the authors find that most reviews on most platforms exhibit a high degree of polarity, but the platforms vary in the degree of polarity on the basis of how selective customers are in reviewing products on the platform. Using cross-platform and multimethod analyses, including secondary data, experiments, and survey data, the authors empirically confirm polarity self-selection, described as the higher tendency of consumers with extreme evaluations to provide a review as an important driver of the polarity of review distributions. In addition, they describe and demonstrate that polarity self-selection and the polarity of the review distribution reduce the informativeness of online reviews.

Keywords

imbalance, online reviews, polarity, self-selection, user-generated content

Online supplement: https://doi.org/10.1177/0022243720941832

Consumer online reviews have become an integral part of the consumers' decision-making process. A recent study found that online reviews influence purchase decisions for 93% of consumers (Kaemingk 2019), and 91% of consumers trust online reviews as much as personal recommendations (Igniyte 2019). Online reviews have also been shown to have an economic impact (e.g., Chevalier and Mayzlin 2006; Dellarocas, Zhang, and Awad 2007; Liu 2006; Moe and Trusov 2011).

One common finding in the study of online reviews has been that reviews have a mass at the positive end of the rating scale, with a few reviews in the midrange and some reviews at the negative end of the scale (Hu, Pavlou, and Zhang 2017; Moe, Netzer, and Schweidel 2017).[1] Indeed, analyzing all consumer reviews of 24 product categories of the e-commerce retailer Amazon, we find that the aggregate distributions of reviews in all 24 categories shows this pattern of polarity in the distribution of reviews.[2] This finding is surprising given that online reviews represent crowdsourcing of preferences and experiences of a large body of heterogeneous consumers, which often have a normal distribution (Hu, Zhang, and Pavlou 2009).

The tendency to observe primarily positive reviews has fueled the debate on how informative consumer reviews actually are (Fritz 2016; Hickey 2015) and whether these consumer reviews mirror "true" product[3] quality (De Langhe, Fernbach, and Lichtenstein 2015). Survey results show that consumers seem to react to the polarity in the distribution of reviews: 92% of consumers say they will use a local business only if it has an average rating of at least four of five stars (Saleh 2015), which indicates that the average rating acts as a threshold criterion rather than a continuous measure. Thus, the common pattern of online reviews may signal a mismatch between consumers' true preferences and those exhibited in online reviews, potentially hindering the usefulness of these reviews.

We describe the common pattern observed in the distributions of online reviews along two dimensions: polarity and imbalance. Specifically, we define "polarity" as the proportion

[1] Herein, we use the term "online reviews" to refer to numerical ratings consumers provide for products or services. Thus, we use "reviews" and "ratings" interchangeably in the remainder of the article.

[2] See Web Appendix 1.

[3] Throughout the article we use "product" to refer to a product, service, or experience.

Verena Schoenmueller is Assistant Professor, Bocconi University, Italy (email: verena.schoenmueller@unibocconi.it). Oded Netzer is Professor of Business, Columbia Business School, Columbia University, USA (email: onetzer@gsb.columbia.edu). Florian Stahl is Professor of Marketing, University of Mannheim, Germany (email: florian.stahl@uni-mannheim.de).

of reviews that are at the extremes of the scale and "positive imbalance" as proportion of positive (vs. negative) reviews. Polarity thus captures how extreme the distribution of reviews is. Positive imbalance indicates the skewness of the distribution toward the positive side of the scale.

Our aim is to explore the polarity and imbalance of online reviews across platforms, its antecedents, and its downstream consequences. Specifically, the objective of this research is threefold: (1) to investigate how prevalent and robust the polarity and imbalance of the distribution of reviews is across platforms, (2) to analyze the role of polarity self-selection (consumers with more extreme opinions are more likely to write reviews) in explaining the heterogeneity in the distribution of online reviews across platforms, and (3) to explore the possible downstream consequences of polarity self-selection.

Although the polarity of review distributions has been widely acknowledged as the predominant underlying distribution of online reviews (e.g., Hu, Pavlou, and Zhang 2017; Li and Hitt 2008), it is unclear how prevalent the polarity of review distributions is. The majority of previous academic studies on consumer reviews have relied on data from Amazon. In fact, 30 out of 64 papers[4] that investigate numerical rating scales summarized in recent meta-analyses use Amazon reviews (Babić Rosario et al. 2016; Floyd et al. 2014; You, Vadakkepatt, and Joshi 2015). Online reviews on Amazon are indeed characterized by a high polarity and a positive imbalance. Thus, the apparent prevalence of the polarity of review distributions in academic research may be driven by an availability bias, focusing on the prevalence of Amazon reviews. In addition, because the majority of the studies have investigated either a single or a couple of platforms, these studies were not able to explore the systematic variation in the distribution of reviews across review platforms.

To investigate the heterogeneity in the review distributions, we have compiled an expensive data set of over 280 million online reviews generated by more than 24 million reviewers from 25 platforms (e.g., Amazon, Yelp, Expedia), covering more than 2 million products and services and reflecting different types of platforms (e.g., e-commerce, review and comparison sites) and various product/service categories (e.g., books, beers, hotels, restaurants). We find that, while the most dominant distribution of online reviews across platforms and product categories is indeed characterized by high degree of polarity and positive imbalance, online reviews of several platforms and product categories are less polar and positively imbalanced. Moreover, we find that the distribution of reviews of the same product can differ across platforms. This raises the question, what drives the variation in the review distributions across platforms? Using a hierarchical model, we investigate the relationship between the distribution of reviews across platforms and different characteristics of the platforms such as the products reviewed, the platform's business model, the rating scale, and how often people review on the platform. We find that platforms on which people review a large number of products exhibit lower polarity relative to platforms on which people elect to review only selected products. We further find that several other platform characteristics can explain the variation in polarity and imbalance across platforms. Importantly, even controlling for a host of platform characteristics, the frequency in which reviewers review on the platform is a robust driver in explaining the variation in polarity across platforms. We also find a relationship between frequency of reviewing and positive imbalance, though it is less robust than the relationship with polarity.

Accordingly, we subsequently investigate the selectivity in which consumers make the effort to review only products with which they are very satisfied or unsatisfied, which we call "polarity self-selection" (Hu, Pavlou, and Zhang 2017). We use a multimethod approach including secondary data analyses, experiments, and surveys to consistently demonstrate that the number of reviews the reviewer has written on the platform can serve as a managerially relevant, and easy to collect proxy for polarity self-selection. Specifically, we find that reviewers with a higher ratio of products reviewed to products purchased (lower self-selection) exhibit less polar and more balanced distributions of reviews. We further establish polarity self-selection in an experimental setting by manipulating polarity self-selection experimentally while holding all other factors constant. We find that consumers who were asked to review the last experienced product (no polarity self-selection) provided less polar reviews compared with consumers who *selected* the product they wish to review out of all products they have experienced.

Finally, we investigate the downstream consequences of polarity self-selection for key metrics such as sales, objective quality, and review usefulness. We show that the greater the polarity self-selection, the lower the relationship between the average rating of online reviews and downstream behaviors. This result may provide a first explanation for the inconclusive results in previous studies regarding the relationship between the average rating and sales (e.g., Babić Rosario et al. 2016; You, Vadakkepatt, and Joshi 2015) as well as between average ratings and objective quality (De Langhe, Fernbach, and Lichtenstein 2015).

The rest of the article is organized as follows: First, we relate our work to previous research on online and offline word of mouth (WOM) and possible self-selection in generating WOM. Next, we describe the large-scale data set of online reviews that we compiled, including over 280 million reviews. The main body of the paper consists of three sections (see Figure 1). In the first section, we explore the robustness of polarity and imbalance across platforms and the role of polarity self-selection in explaining the variation across platforms. In the second section, we investigate the mechanisms underlying polarity self-selection and the polarity and imbalance of online reviews. In the third section, we investigate how the polarity and imbalance of online reviews can affect the informativeness of these reviews. We conclude with a discussion of our

[4] This includes journal publications and conference proceedings but excludes working papers and unpublished (at the time) dissertations.

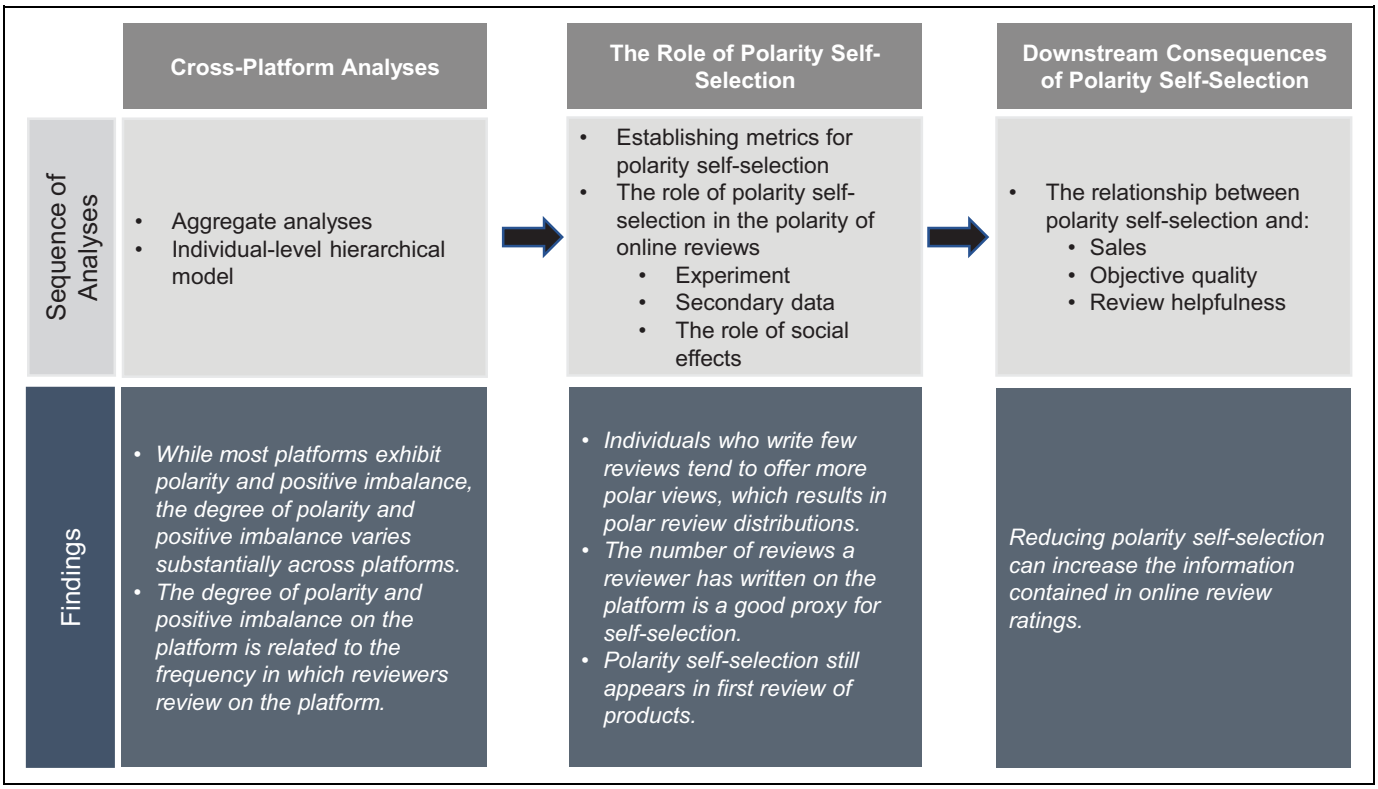

Figure 1. Road map for investigating the polarity and imbalance of online review distributions.

findings, implications for consumers and online review platforms, and an outlook toward future research.

The Polarity of Online and Offline WOM

Our research builds on and extends the findings of previous research that documented the polarity of the distribution of online reviews (e.g., Dellarocas, Gao, and Narayan 2010; Feng et al. 2012; Godes and Silva 2012; Hu, Pavlou, and Zhang 2017; Zervas, Proserpio, and Byers 2015), offline WOM (e.g., East, Hammond, and Wright 2007; Naylor and Kleiser 2000) and consumer satisfaction (e.g., Danaher and Haddrell 1996; Peterson and Wilson 1992). Although prior research has documented the presence of polarity and imbalance in online reviews, it has neither investigated their robustness nor the possible reasons for their variation across platforms.

Self-selection has been suggested as a potential driver of the polarity of review distributions (Li and Hitt 2008). The primary forms of self-selection discussed in the literature are purchase self-selection (Hu, Pavlou, and Zhang 2017; Kramer 2007), intertemporal self-selection (Li and Hitt 2008; Moe and Schweidel 2012), and polarity self-selection (Hu, Pavlou, and Zhang 2017). In the context of online reviews, the most discussed form of self-selection is purchase self-selection—that is, consumers who are a priori more likely to be satisfied with a product are also more likely to purchase it, and thus, the initial group of potential reviewers might already be more positive about the product than the general population (Dalvi, Kumar, and Pang 2013; Kramer 2007). Purchase self-selection has also been discussed in the satisfaction literature, suggesting that, on average, consumers are often satisfied with the product they purchase (e.g., Anderson and Fornell 2000; Lebow 1982; Mittal and Kamakura 2001; Mittal and Lassar 1998; Peterson and Wilson 1992). We note, however, that it is not clear whether one could call purchase self-selection in the context of online reviews a "bias" per se, as consumers who intend to buy the product may be interested in the preferences of only the self-selected group of consumers who were interested enough in the product to purchase it. Assuming that most consumers who reviewed a product bought it (a few exceptions might include fake reviews or incentivized reviews), purchase self-selection alone cannot explain the variation in the polarity of the review distributions across platforms/products/reviewers/reviews as it is likely to affect all of the reviews. That being said, purchase self-selection is likely to play a role in the observed polarity and positive imbalance of online reviews relative to the preferences of the entire consumer universe.

Intertemporal self-selection arises when consumers at different times in the product or consumer life cycle elect to review products. For example, Li and Hitt (2008) demonstrate that earlier reviews in the product life cycle tend to be extreme and positive due to self-selection of the type of reviewers (early vs. late adopters), giving rise to a polar distribution early on in the product life cycle. Another form of intertemporal self-selection is due to social influence. Seeing previous reviews can influence one's motivation to review as well as the actual

review provided (Godes and Silva 2012; Moe and Schweidel 2012; Moe and Trusov 2011; Schlosser 2005).

In addition to purchase self-selection, Hu, Pavlou, and Zhang (2017) and Dellarocas, Gao, and Narayan (2010) also discuss self-selection due to consumers' greater propensity to review products, with which they had either extremely good or bad experiences (polarity self-selection). The tendency to weigh negative and positive experiences more strongly is rooted in social psychology (Skowronski and Carlston 1987) and applied in the context of offline and online WOM (e.g., Berger 2014; Schlosser 2005). It has been suggested that extreme cues are perceived as less ambiguous and more diagnostic, and thus they receive heightened attention (Gershoff, Mukherjee, and Mukhopadhyay 2003). The WOM literature suggests mixed results with respect to the likelihood of satisfied and dissatisfied consumers to spread WOM. Some suggest that dissatisfied consumers are more likely to spread WOM (e.g., Heskett, Sasser, and Schlesinger 1997; Silverman 1997), whereas others find a higher likelihood for satisfied consumers (e.g., East, Hammond, and Wright 2007; Engel, Kegerreis, and Blackwell 1969). Anderson (1998) reports a U-shaped frequency of offline WOM for satisfied and dissatisfied consumers. Online reviews, in contrast, have often been characterized as being polar and positively imbalanced resulting in a J-shaped distribution (Moe, Netzer, and Schweidel 2017). One could rationalize this discrepancy between the pattern of offline satisfaction and online WOM distribution with the following three findings: (1) writing online reviews is generally more effortful compared with sharing offline WOM, and thus consumers may be less likely to exert the effort to report mediocre experiences (King, Racherla, and Bush 2014); (2) WOM in the online environment is often transmitted over weaker ties, and individuals tend to be reluctant to transmit negative information to weaker ties (Zhang, Feick, and Mittal 2014); and (3) while offline WOM is often aimed at only one person or a small group of people, online reviews are accessible by a considerably larger audience (Dellarocas 2003). Barasch and Berger (2014) show that when communicating with multiple individuals, people are less likely to share negative information to avoid sharing content that makes them look bad. We build on that literature and demonstrate how polarity self-selection can be used to explain the variation in the review distribution across platforms. We find that polar review distributions imbalanced to the positive side of the scale (J-shaped distribution) exist across multiple products and platforms but exhibit variation that can be meaningfully explained by the degree of polarity self-selection.

In addition to polarity self-selection, review fraud (e.g., Anderson and Simester 2014; Luca and Zervas 2016; Mayzlin, Dover, and Chevalier 2014) has been proposed to explain the polarity and imbalance of the review distribution. For example, Luca and Zervas (2016) find that fake reviews on Yelp exhibit a higher degree of polarity relative to other reviews. Similarly, Mayzlin, Dover, and Chevalier (2014) show that hotels neighboring a hotel with a high incentive to post fake reviews are more likely to have one- and two-star (negative) reviews or five-star (positive) reviews, with the effect being more consistent for negative reviews. Finally, Anderson and Simester (2014) find that, relative to verified reviews, unverified reviews are negatively imbalanced. Taken together, this stream of research suggests that review fraud can possibly lower the positive imbalance due to a larger number of negative reviews. To account for review fraud, we include a measure of the platform's mechanism to verify reviews in our cross-platform analyses.

The polarity and imbalance of the distribution of consumer evaluations can also arise from the format of the scale used to elicit the evaluations. The satisfaction and psychometric literature indicate that while scale modifications (e.g., question framing, number of scale points, multi-item scales, scale wording) can affect the resulting distribution (e.g., Danaher and Haddrell 1996; Moors, Kieruj, and Vermunt 2014; Weijters, Cabooter, and Schillewaert 2010), scale modifications alone cannot eliminate polarity and imbalance of response distributions (Peterson and Wilson 1992). To account for possible effects of scale characteristics on the scale distribution, we include scale characteristics in our cross-platform analyses.[5]

Review Distributions Across Platforms

To investigate how robust and generalizable the polarity and imbalance of the review distribution is across platforms and to explain the variation across platforms, we collected a comprehensive data set with more than 24 million reviewers, 2 million products, and a total of over 280 million online reviews. We collected reviews from 25 platforms, including Amazon, a European online retailer, Epinions, RateBeer, MovieLens, The Internet Movie Database (IMDb), Rotten Tomatoes, Yahoo! Movies, Fandango, Edmunds, Twitter, Yahoo! Songs, Netflix, Trustpilot, Metacritic, Goodreads, Yelp, TripAdvisor, Expedia, Airbnb, BlaBlaCar, Google Restaurant reviews, Booking.com, yourXpert, and Frag-Mutti. We selected all platforms with respect to their dominant position in their respective industries (according to Alexa Rank and Google Trends). Table 1 provides an overview of the 25 platforms and the number of products and reviews that we have sampled.

As can be seen in Table 1, the platforms are quite heterogeneous along several dimensions. For example, platforms vary with respect to their business model (selling products/services, collecting transaction fees, or information platforms primarily collecting revenue from advertising), product category, or their approach to collecting and verifying reviews. As we discuss and demonstrate subsequently, these factors could be related to the degree of polarity and imbalance of the review distributions on these platforms. Using the cross-platform data set we assembled, we examine how robust the polarity and imbalance of online reviews are across product categories and online platforms. In addition, we investigate different platform

[5] We also examined, in a lab setting, the effect of variations of the commonly used scales and scale wordings on the resulting review distribution. We did not find a significant impact of these scale characteristics. Details of the analyses are available from the authors.

Table 1. Platform Characteristics, Polarity and Positive Imbalance Across the 25 Platforms Used in this Article.[a]

Platform	Product Category	Type of Business Model	Age of Platform (Years)	Reviewer Social Network	Reviewer Recognition	Verified Reviews	Popularity Ranking[b]	Response to Reviews	Scale Points	Polarity[c]	Imbalance[c]	# Products in Our Sample	# Reviews in Our Sample	Average # Reviews/ Reviewer (Mdn)
yourXpert	Products and services	Transaction fee	6	No	No	Yes	577,484	Yes	5	87%	95%	78	4,733	1 (1)
Trust Pilot	Products and services	Information platform	12	No	No	No	477	Yes	5	86%	76%	92	202,242	3 (1)[i]
Frag-Mutti	Products and services	Information platform	16	No	No	No	59,316	Yes	5	85%	83%	1,811	26,224	21 (1)
BlaBlaCar	Travel/restaurants	Transaction fee	13	No	Yes	Yes	17,247	Yes	5	84%	99%	1,075	52,456	29 (13)[i]
Airbnb	Travel/restaurants	Transaction fee	11	Yes	No	Yes	230	Yes	5	72%	97%	1,404	48,571	4 (2)[i]
Amazon[d]	Products and services	Selling products/services	25	Yes	Yes	Yes	13	Yes	5	68%	85%	2,008,781	68,700,502	4 (1)
Google Restaurants	Travel/restaurants	Information platform	21	No	Yes	No	1	Yes	5	65%	92%	744	242,134	38 (13)
Fandango	Entertainment	Information platform	19	No	No	No	2,081	No	10	65%	74%	105	96,540	—[j]
Online Retailer	Products and services	Selling Products/services	7	No	Yes	Yes	6,240	Yes	5	65%	89%	8,305	555,974	2 (1)
Edmunds	Products and services	Information Platform	24	No	No	No	1,973	Yes	5	63%	94%	4,784	179,640	1 (1)
Booking.com	Travel/restaurants	Transaction fee	23	No	No	Yes	63	Yes	10	60%	97%	1,492	515,738	7 (3)
Epinions	Products and services	Information platform	20	Yes	Yes	No	130,034	No	5	56%	73%	11,481	147,149	12 (2)
Yahoo! Songs (Launchcast)	Entertainment	Transaction fee	18	Yes	No	No	—	No	5	56%	48%	1,000	311,704	—[j]
Expedia	Travel/restaurants	Transaction fee	23	No	No	Yes	458	Yes	5	52%	84%	4,990	265,145	—[j]
Yelp	Travel/restaurants	Information platform	15	Yes	Yes	Yes	186	Yes	5	51%	69%	63,154	4,666,385	27 (7)
Yahoo! Movies[e]	Entertainment	Information platform	21	No	No	No	11	No	13	50%	83%	3,382	205,809	29 (17)
TripAdvisor	Travel/restaurants	Transaction fee	19	Yes	Yes	No	236	Yes	5	44%	75%	6,475	1,100,156	46 (11)[i]
Metacritic	Entertainment	Information platform	18	No	No	No	2,039	No	11	42%	75%	824	45,803	31 (5)[i]
Rotten Tomatoes	Entertainment	Information platform	21	No	Yes	No	599	No	10	41%	69%	303	72,454	460 (107)[i]
Goodreads	Products & services	Information platform	13	Yes	Yes	No	311	No	5	38%	90%	888	60,917,897	575 (226)[i]
IMDb[f]	Entertainment	Information platform	23	No	Yes	No	54	No	10	36%	69%	27,241	21,181,881	367 (32)[i]
Netflix	Entertainment	Selling products/services	22	No	No	No	21	No	5[h]	29%	61%	17,770	100,481,301	209 (96)
Twitter Movies[g]	Entertainment	Information platform	13	Yes	No	No	35	No	10	24%	85%	7,484	516,199	12 (2)
MovieLens	Entertainment	Information platform	22	No	No	No	148,026	No	10	23%	71%	22,795	24,367,613	94 (29)
RateBeer	Products and services	Information platform	19	Yes	Yes	No	45,146	No	20	9%	82%	28,521	1,503,127	69 (4)

[a]Web Appendix 3 provides an overview of our data sources used and links to a data repository.
[b]The rankings of website traffic were gathered via https://www.alexa.com/siteinfo. Rankings are available only for the entire platform as opposed to the specific sections that are included in our data set.
[c]Only products with more than five reviews were used in order to calculate a stable distribution.
[d]We examined 24 Amazon product categories.
[e]We use reviews of the former Yahoo! Movies platform, when Yahoo! still generated its own reviews.
[f]The original data set gives the proportion of reviews in each rating bracket rounded up to the nearest 5%.
[g]Movie ratings for Twitter are ratings from users given on IMDb and then tweeted.
[h]Since the time of our data collection, Netflix moved to a two-point scale.
[i]For these platforms, we have no access to the number of reviews per reviewer of the entire sample. We thus approximate the number of reviews per reviewer by drawing a random sample of reviewers and the reviews they have written.
[j]Due to data limitations, we could not access the number of reviews per reviewer for these platforms.

characteristics that can possibly explain the variation in the review distribution across platforms.

We start by defining the measures of the review distribution for the most common five-point scale (68% of the platforms in our data set), for polarity and positive imbalance.[6,7]

$$\text{Polarity} = \frac{\text{Number(one- and five-star ratings)}}{\text{Number of ratings}} \quad (1)$$

$$\text{Positive imbalance} = \frac{\text{Number(four- and five-star ratings)}}{\text{Number(one-, two-, four-, and five-star ratings)}}. \quad (2)$$

According to Equation 1, for a five-point scale, a polarity measure above 40% implies a polar distribution, whereas a polarity measure below 40% implies a nonpolar distribution. Equation 2 provides a measure of the skewness of the distribution to the positive side of the scale such that an imbalance measure above 50% means that there are more positive reviews and below 50% indicates a majority of negative reviews. Thus, our measure of imbalance captures the positive imbalance of the reviews. When relating our measure of positive imbalance to different factors (e.g., number of reviews per reviewer), a positive (negative) effect would mean that the factor leads to more positive (negative) reviews.

To make the analysis comparable across platforms with different scale lengths, we rescaled the scales of platforms with a scale longer than five points before applying Equations 1 and 2 such that polarity is defined as the extreme 20% of the scale on each side of the scale and imbalance as the 60%+ positive scale points. For scales divisible by five, this rescaling is straightforward. For scales not divisible by five, 20% or 80% of the scale does not lead to an integer scale point. Thus, one could scale to the closet scale point to the 20% or 80% cutoff. In addition, for such scales we recommended testing the robustness of polarity and imbalances for the scale points on the two sides of the 20% or 80% cut off. For example, for the two scales in our data set that were nondivisible by five (Yahoo Movies! has a 13-point scale and Metacritic has an 11-point scale), we define polarity based on the number of reviews with 1–2 and 12–13 stars as well as 0–1 and 9–10 stars, respectively, and test the robustness of the results for a polarity definition based on 1–3 and 11–13 stars as well as 0–2 and 8–10 stars, respectively. For all scales used in this article, we provide the scale transformation to calculate polarity and positive imbalance in Web Appendix 2. We also test the robustness of our definition of polarity and positive imbalance using only the extreme scale points to measure polarity in longer scales (e.g., one and ten in a ten-point scale).

[6] In the subsequent analyses we use log(1 + polarity) and log(1 + positive imbalance).

[7] Our measure of positive imbalance is related to the measure used by Fisher, Newman, and Dhar (2018), who used the difference between four- and five-star reviews and one- and two- star reviews.

Variation of online review distributions across online platforms. While review distributions with a high degree of polarity have been documented to be the prevalent distribution of online reviews both in the academic research (e.g., Chevalier and Mayzlin 2006; Hu, Pavlou, and Zhang 2017; Kramer 2007) and in popular press (Hickey 2015; Wolff-Mann 2016), several studies have found distributions with a lower degree of polarity on platforms such as Yelp, MovieLens, Netflix, and Yahoo! Songs (e.g., Dalvi, Kumar, and Pang 2013; Luca and Zervas 2016; Marlin et al. 2012). To investigate the generalizability of polarity and imbalance in the review distributions across platforms, we compare the review distributions across the 25 platforms in our data set. As shown in Table 1 and Figure 2, there exists significant heterogeneity across platforms with respect to the prevalence of polarity. While more than two-thirds of the platforms have polar distributions (e.g., Amazon, Google Restaurant reviews, BlaBlaCar, or Airbnb), platforms such as RateBeer, MovieLens, or Netflix are characterized by an average polarity below 30%. While there is substantial heterogeneity in the polarity of the distribution across platforms, we find that almost all platforms are imbalanced toward positive reviews. That being said, some platforms, such as the sharing economy platforms (Airbnb and BlaBlaCar), exhibit stronger positive imbalance, whereas at the other extreme Yahoo! Songs exhibits a balanced distribution.

In Table 1 we also compare the different platforms on several platform characteristics: (1) average number of reviews written per reviewer, (2) business model (selling products/services, charging a fee per transaction, providing information), (3) product category (products and services, entertainment, and travel/restaurants), (4) length of rating scale, (5) existence of a social network among reviewers, (6) existence of reviewer recognition, (7) flagging or requiring verified reviews, (8) platform popularity ranking as measured by web traffic, and (9) opportunity for sellers to respond to reviews. A cursory analysis of Table 1 reveals several interesting patterns. First concerning the business model, the review distributions of platforms either selling products or receiving a fee from each transaction (e.g., Amazon, Expedia) are more polar and positively imbalanced relative to information platforms (e.g., MovieLens [an academic movie recommender system]). This may suggest that more commercial platforms have an incentive to showcase positive reviews to entice consumers to buy (Hickey 2015). While primarily anecdotal, given that we have only two two-sided platforms in our data set (Airbnb and BlaBlaCar), these platforms exhibit high degree of polarity and positive imbalance, which could be explained by reciprocity in review behavior (Dellarocas and Wood 2008). Polarity and positive imbalance also seem to vary by product category. However, considerable heterogeneity also exists across platforms within the same product category (e.g., movies on MovieLens vs. Yahoo! Movies).

Platforms that use longer scales (e.g., RateBeer, MovieLens, and IMDb) exhibit a lower degree of polarity and positive imbalance relative to platforms that use a five-point

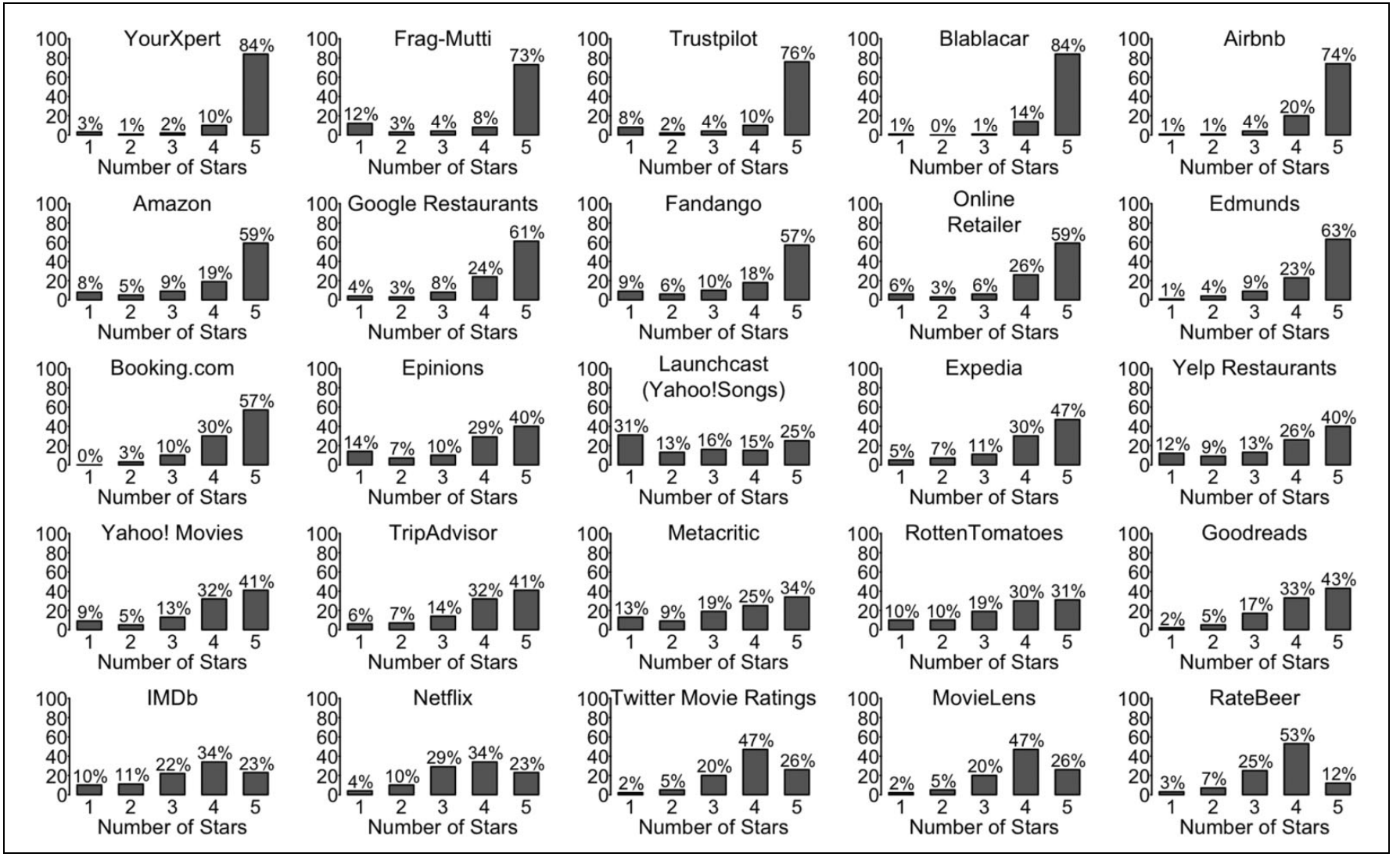

Figure 2. Distribution of online reviews across the online platforms.

scale.[8] This pattern may suggest that the five-point scale used in most online platforms leads to right-censored review patterns relative to longer scales. Having a social network among reviewers as a feature of the platform (e.g., Yelp) or other forms of reviewer recognition might stimulate the activity and the engagement of reviewers with the platform and thus might reduce polarity. Indeed, we see that many of the platforms that have such social networks or recognition (e.g., Yelp, Goodreads, Tripadvisor, RateBeer) also exhibit lower review polarization. Platforms further differ on whether reviewers or purchase occasions are verified. For example, Expedia allows for reviews only from customers who purchased the product, and Amazon marks reviews of customers who purchased the product on Amazon as "verified." Given that verification is likely to reduce the degree of review fraud, and fraudulent reviews have been shown to exhibit higher polarity (Mayzlin, Dover and Chevalier 2014), we may expect to see lower polarity distributions for platforms with verified reviews. However, Table 1 seems to suggest an opposite pattern. Another characteristic of platforms is their general popularity (number of people who visit the platform daily). On the one hand, a more popular platform might increase the engagement of the reviewers on the platform leading to lower self-selection and, thus, lower polarity. On the other hand, popularity might attract a higher ratio of one-time or infrequent reviewers, leading to higher polarity. We also investigate the opportunity for sellers to respond to reviews of reviewers. The ability of sellers to respond may deter mediocre or negative reviewers from posting a review, leading to higher polarity and positive imbalance.

Finally, one of the most important variables in our analysis is the number of reviews a reviewer writes on the platform, which, as we demonstrate in the next section, is a good proxy for polarity self-selection. The rationale is that individuals who review a larger fraction of products they purchased/experienced are less selective in the products they review relative to individuals who review only one or a couple of products.[9] Table 1 provides first indications that, in line with the polarity self-selection account, both polarity and positive imbalance decrease when the number of reviews per reviewer increases.

To more systematically assess the relationship between the different platform characteristics and polarity and positive imbalance, we use two analyses. First, taking a platform as a unit of analysis, we regress the measures of polarity and positive imbalance on the different platform characteristics. Given the limited number of platforms, and to preserve degrees of freedom rather than including all platform characteristics, we regress polarity and positive imbalance on each platform characteristic, one at a time. In addition, we regress polarity and positive imbalance on each platform characteristic together with our proxy for polarity self-selection (number of reviews per reviewer). Given the limited number of observations (platforms) that analysis should be taken as primarily descriptive in nature. Accordingly, in a second analysis we conduct a meta-analytic approach, using a sample of individual reviews across platforms as the unit of analysis and "stacking" reviews from all platforms together (leading to $N = 17{,}200$) to analyze the relationship between platform and reviewer characteristics and polarity. We do so by estimating an individual-level hierarchical model on the stacked data across platforms.[10]

Aggregate cross-platform analysis. The regression of polarity and positive imbalance on each platform characteristic one at a time (see Table 2) shows that polarity self-selection (log number of reviews per reviewer) explains a large portion of the variance ($R^2 = 39\%$) in the polarity across platforms as well as in the degree of positive imbalance ($R^2 = 31\%$). For polarity, we also find that whether companies can respond to a review, the number of scale points, and product category can explain a substantial proportion of variance across platforms, while for positive imbalance we find that the product category, whether companies can respond to a review, whether reviews on the platform are verified, and the business model can explain a substantial proportion of variance.[11]

To more closely assess the marginal effect of each platform characteristic above and beyond polarity self-selection, we ran a regression of each characteristic together with the polarity self-selection measure on polarity and positive imbalance. We find not only that polarity self-selection explains a substantial portion of the variance of polarity and positive imbalance but also that, controlling for the other characteristics, polarity self-selection is always significantly related to polarity and positive imbalance of the distribution except for the seller's opportunity to answer to reviews (for details, see Web Appendix 4).

Individual level hierarchical model. The cross-platform analysis at the platform level provided first evidence with respect to the factors that can lead to polarity and positive imbalance in reviews and confirmed that the number of reviews per reviewer (polarity self-selection), even controlling for other potential drivers, shows a significant relationship with polarity and positive imbalance of the review distribution. However, due to the limited number of platforms this analysis is primarily directional. To further examine these relationships, while overcoming the small sample size that arises from the analysis at the platform level, we extend our analysis to the individual review, as opposed to the platform level, thus increasing the number of observations substantially. Specifically, we "stack" the reviews across platforms and use a multilevel model with a platform random effect. This analysis enables us to investigate both

[8] Recall that in calculating polarity and positive imbalance we rescale these longer scales to the corresponding five-point scale.

[9] One possible concern with this measure is that it captures the number of reviews one has written without considering the number of products purchased. In a subsequent analysis we show that the number of reviews is a good proxy for the ratio of reviews to products purchased (self-selection).

[10] In this analysis we cannot use positive imbalance as our dependent variable because positive imbalance cannot be measured at the individual review level.

[11] Replacing the average number of reviews per reviewer with the median number of reviews per reviewer leads to similar results.

Table 2. Variance Explained of Cross-Platform Polarity and Positive Imbalance by Platform Characteristics.

	Polarity	R^2	Positive Imbalance	R^2	N
log(average # reviews per reviewer)	−.050 (.014)***	.393	−.018 (.006)***	.305	22
Age of platform	−.010 (.006)*	.137	−.004 (.002)*	.130	22
Business Model (Reference: Selling Products or Services)					
Transaction fee	.098 (.105)	.162	.078 (.040)*	.261	22
Information platform	−.045 (.091)		.007 (.035)		
Product Category (Reference: Products/Services)					
Travel/restaurants	.014 (.068)	.324	.015 (.028)	.332	22
Entertainment	−.171 (.065)**		−.067 (.026)**		
Scale points	−.189 (.053)***	.392	−.028 (.027)	.052	22
Network among reviewers	−.087 (.065)	.083	−.001 (.028)	.000	22
Reviewer recognition	−.035 (.065)	.014	−.012 (.027)	.010	22
Verified reviews	.157 (.060)**	.254	.063 (.025)**	.242	22
Popularity ranking	.004 (.009)	.009	.001 (.004)	.003	22
Seller ability to respond to reviews	.229 (.041)***	.614	.065 (.023)***	.292	22

*$p < .1$.
**$p < .05$.
***$p < .01$.
Notes: Each row in this table is a separate regression. SEs in parentheses. We did not include three platforms for which we had no access to the number of reviews per reviewer.

platform and reviewer characteristics with greater statistical power, and controlling for all characteristics simultaneously. We fit the following hierarchical model:

$$Y_{ij} = \alpha_j + \beta x_{ij} + \varepsilon_{1ij},$$

$$\alpha_j = \gamma \times z_j + \varepsilon_{2j},$$

where Y_{ij} is the polarity of the review i posted on platform j (1 if the rating was one or five, and 0 otherwise). x_{ij} is the number of reviews a reviewer has written on a platform and the number of reviews of the product reviewed. The higher-level equation relates polarity to the platform random effect (α_j) and the review specific covariates (log number of reviews per reviewer and log number reviews of product). The lower-level equation relates the platform random effects to the set of platform characteristics (z_j) described in the aggregate level analysis. To ensure that all platforms weigh equally in our analysis, we randomly sampled 1,000 reviews per platform[12] and ensure that each review belongs to a unique reviewer. As can be seen in Table 3, the random-effect multilevel model reveals a strong and significant relationship between polarity self-selection and polarity. We also find that platforms that sell products and services exhibit significantly stronger polarity relative to other types of platforms. In addition, platforms with longer scales exhibit significantly lower polarity relative to other platforms. When we control for other platform characteristics, we find that the presence of review verification and sellers' ability to respond to reviews are no longer significantly associated with review polarity. Overall, the results of the individual-level hierarchical model confirm our findings from the aggregate cross-platform analysis with greater statistical power. To more directly measure the relationship between the number of reviews per reviewer and polarity across platforms, while fully accounting for the variation across platforms, we also run a platform fixed effect model with the number of reviews of a reviewer as independent variable. Again, we find a significant relationship between the number of reviews a reviewer has written and the polarity of the reviews: ($\beta_{\#\text{ of reviews per reviewer}} = -.036$ [.003], $p<.01$).

In summary, we find that while many platforms exhibit polarity and positive imbalance, this is not the case for all platforms, suggesting that the focus of past research on few platforms such as Amazon may have created a distorted belief about the prevalence and the degree of polarity and positive imbalance of online review distributions. In addition, we find that the number of reviews per reviewer, as a proxy for self-selection, can explain a large portion of the variation in the polarity of the review distributions across platforms.

Within-platform analysis: Yelp reviews. Our cross-platform analysis has established that the number of reviews a reviewer wrote on the platform is a strong and robust predictor of the polarity of the review distribution. However, because platforms simultaneously differ with respect to multiple characteristics, we aim to further investigate polarity self-selection and its impact on the polarity and positive imbalance of the review distribution

[12] For five platforms, we could only access a smaller sample. See Web Appendix 5 for details.

Table 3. A Random-Effect Hierarchical Model of Polarity Self-Selection and Platform Characteristics on Polarity.[a]

Polarity Self-Selection: Log(Number of Reviews per Reviewer)	−.163 (.013)***
Log(number of reviews per product)	−.008 (.011)
Age of platform	−.046 (.027)*
Business Model (Reference: Selling Products or Services)	
Transaction fee	−2.009 (.699)***
Information platform	−.824 (.446)*
Product Category (Reference: Products/Services)	
Travel/restaurants	−.897 (.518)*
Entertainment	.392 (.530)
Scale points	−1.090 (.399)***
Network among reviewers	−.316 (.370)
Reviewer recognition	.190 (.276)
Verified reviews	.493 (.437)
Popularity ranking	−.010 (.039)
Seller ability to respond to reviews	.605 (.507)
Constant	2.864 (.974)***
N	17,200

*$p < .1$.
**$p < .05$.
***$p < .01$.
[a]We also test the robustness of the results to using only the extreme points (for scales longer than five-points) and alternative scale operationalizations for platforms with scales not divisible by five (Metacritic and Yahoo! Movies) in calculating polarity and positive imbalance and find similar results. In addition, we rerun the model using only a sample of 100 reviews per platform and only including platforms with at least 1,000 reviews and find similar results. See Web Appendix 6.
Notes: SEs in parentheses. This analysis included 21 platforms. In addition to the three platforms for which we had no access to the number of reviews per reviewer on these platforms, we also had to exclude the Airbnb platform because we did not have access to online ratings at the reviewer level on Airbnb.

using a within-platform analysis, thus, holding all platform characteristics constant.

To investigate how polarity and positive imbalance differ based on the reviewer frequency of reviews we analyze the review distribution of Yelp reviewers with varying frequency of reviews. First, to visually depict the relationship between review distribution and frequency of reviews, we split all Yelp reviewers in our data set to the upper ($n_{\text{upper quartile}} = 89{,}096$) and lower ($n_{\text{lower quartile}} = 88{,}947$) quartiles, based on the number of reviews written by a reviewer for a restaurant per month since joining Yelp.[13] We calculate the number of reviews per month by dividing the number of reviews a reviewer has written by the number of months she has been a member of Yelp.[14] Figure 3, Panel A, compares the review distributions between frequent and infrequent reviewers. Consistent with our cross platform analysis, we see that, even within a platform, the distribution of online reviews of frequent reviewers is not polar, whereas the review distribution of infrequent reviewers exhibits a high degree of polarity.

To statistically analyze the relationship between the frequency of reviewing and polarity self-selection, and to go beyond the visual dichotomization of frequency of reviewing in Figure 3, we regress polarity and positive imbalance of all of the reviewer's restaurant ratings on the number of reviews written by a reviewer per month as a covariate. To supplement the cross-platform analysis, and to further examine the effect of the reviewer's network on polarity and positive imbalance, we include the number of followers a user has (a one-way relationship), the number of friends the user has (a two-way relationship) and the number of years a reviewer has been an "Elite" member as covariates. As Table 4 shows, we find that the larger the number of reviews the reviewer wrote per month, the less polar is the distribution of her reviews. However, for positive imbalance we find a positive significant effect, demonstrating that the relationship between positive imbalance and the number of reviews per reviewer is less consistent compared with polarity. Regarding social characteristics of reviewers, we find that the more followers a reviewer has, the less polar and positively imbalanced the review distribution becomes; however, the opposite is true for the number of friends a reviewer has. In addition, we find that Elite members write less polar reviews.

One could argue that frequent and infrequent reviewers visit different restaurants and thus exhibit different distributions. To investigate this alternative explanation, we compare the distributions of frequent and infrequent reviewers within a restaurant. We again use the Yelp data set ($n_{\text{restaurants}} = 36{,}882$, $n_{\text{reviews}} = 3{,}391{,}872$[15]). For each restaurant, we split the reviewers via the lower and upper quartile of the review frequency of each restaurant.[16] We investigate whether polarity and imbalance of the infrequent reviewers is significantly higher compared with those of frequent reviewers. Consistent with polarity self-selection, we find a higher proportion of one- and five-star reviews for infrequent ($m = 56.52\%$) compared with frequent reviewers ($m = 36.10\%$) for the same restaurant ($z = 268.561$, $p < .001$). To investigate the within-restaurant difference, we run a two-tailed sign test on the pairwise differences of the polarity and positive imbalance between the two quartile reviewer groups ("frequent reviewers" vs. "infrequent reviewers") for each restaurant. Of the 36,882 restaurants, 28,717 (79.17%) have a higher proportion of one- and five-star reviews for infrequent reviewers, while only 7,557 (20.83%) restaurants have a higher proportion of one- and five-star reviews for frequent reviewers (608 have the same proportion of one- and five-star reviews for the two groups). A sign test suggests that this difference is significant ($z = 111.106$,

[13] We exclude reviewers with fewer than three restaurant reviews. The result is robust if we include all Yelp reviewers.
[14] We approximate the number of months a reviewer has been on Yelp via the time difference between the date the reviewer joined the platform and the date of her last review.
[15] Because we are doing an analysis within a restaurant, we need to ensure a sufficient number of reviews per restaurant. Accordingly, we use only restaurants with more than 15 reviews and exclude reviewers that have fewer than 3 restaurant reviews. The result is robust if we include all restaurants and reviewers.
[16] We replicate the same analyses for polarity and positive imbalance using a median split and find similar results.

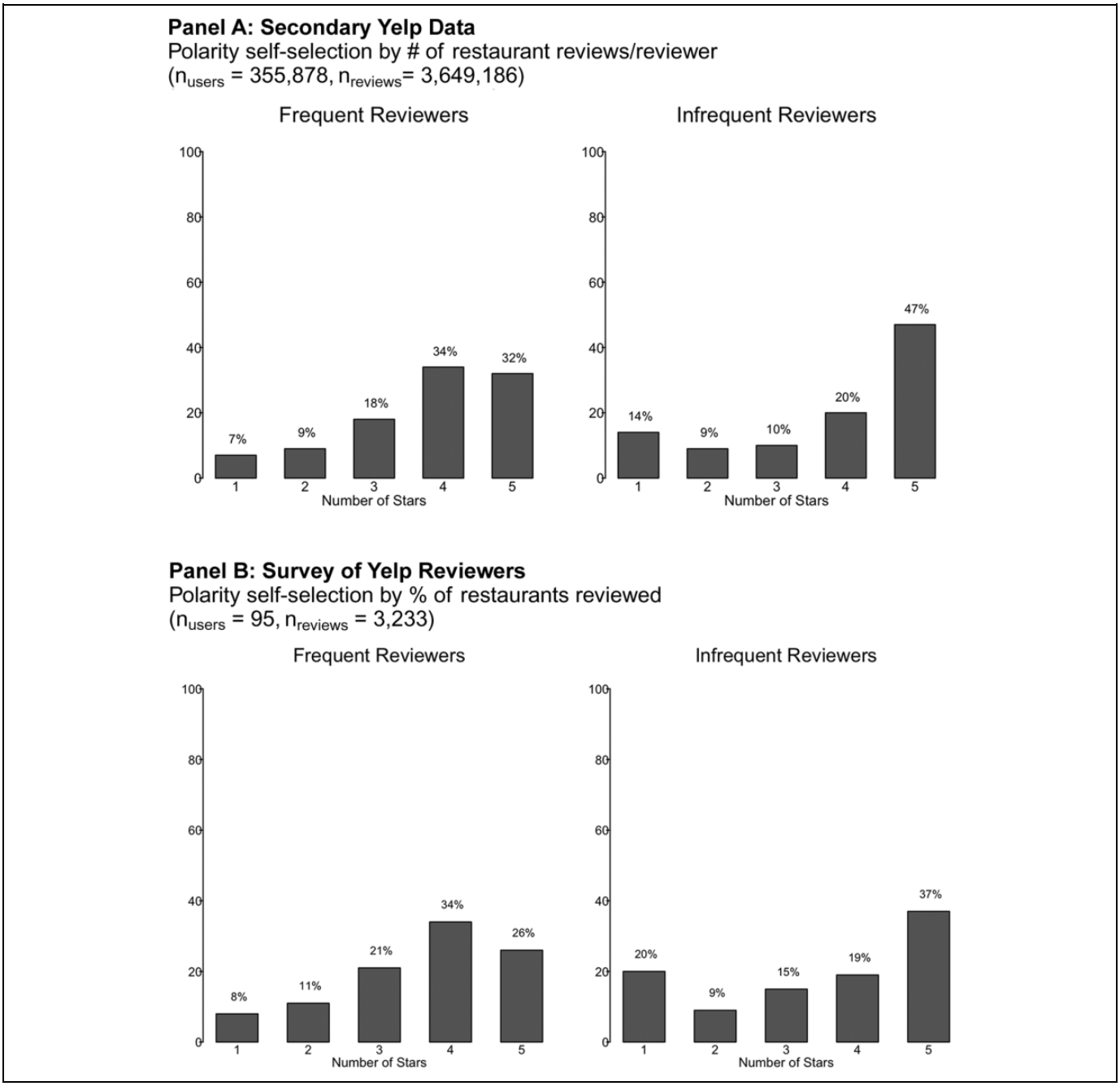

Figure 3. Review distribution of frequent and infrequent Yelp reviewers.

$p < .001$). For positive imbalance, 13,000 restaurants (37.23%) have a higher positive imbalance for infrequent reviewers, while 21,921 (62.77%) have a higher positive imbalance for frequent reviewers (1,940 have the same imbalance of reviews). This difference is significant (sign test: $z = -47.733, p < .001$). Thus, similar to the previous analysis, we find that positive imbalance is stronger for frequent reviewers (lower self-selection).

Overall, this analysis rules out the possibility that polarity self-selection is driven by frequent and infrequent restaurant dwellers visiting different restaurants. Building on the these results, we examine the number of reviews written by a reviewer as a proxy for self-selection and compare it with other possible proxies in the next section. For this investigation, we use experimental and secondary data to reveal the role of polarity self-selection in the prevalence and the degree of the polarity of online reviews.

Polarity Self-Selection and the Distribution of Reviews

Having established the number of reviews per reviewer as a strong predictor of the polarity and positive imbalance of the

Table 4. Polarity Self-Selection: Within Platform-Analysis (Yelp Reviews).[a]

	Polarity	Positive Imbalance
Intercept	.442 (.001)***	.546 (.001)***
Log(number restaurant reviews by reviewer per month)	-.021 (.000)***	.011 (0.000)***
Number of years Elite batch	-.036 (.000)***	.012 (.000)***
Number of followers (in 1,000s)	-.217 (.018)***	-.261 (.016)***
Number of friends (in 1,000s)	.030 (.002)***	.061 (.002)***
R^2	.064	.018
N	355,878	355,589[b]

***$p < .01$.
[a]We exclude reviewers with fewer than three restaurant reviews. The result is robust when we include all reviewers.
[b]For reviewers who only wrote three-star reviews, positive imbalance cannot be calculated.
Notes: SEs in parentheses.

review distribution both across platforms and within a platform, in this section we conducted a survey among Yelp reviewers to examine the validity of the number of reviews a reviewer wrote as proxy for polarity self-selection, as well as two additional proxies: median time between reviews and the standard deviation of time between reviews. In addition, to establish the causal effect of polarity self-selection on the polarity of the review distributions, we use an experimental setting in which we manipulate polarity self-selection directly.

Yelp Reviewers' Survey

In the analysis thus far, we relied on the assumption that more frequent reviewers are less selective in the products they choose to review relative to less frequent reviewers. However, in the context of Yelp, as an example, it is possible that more frequent reviewers are not less selective but, rather, go more frequently to restaurants and are as selective or even more selective in terms of the proportion of restaurants they select to review. To directly measure the degree of polarity self-selection—the proportion of restaurants reviewers chose to review of those they visited—we augmented the Yelp reviews' secondary data with a survey of Yelp reviewers asking them about the frequency of their restaurant visits. We recruited via Amazon Mechanical Turk (MTurk) Yelp reviewers who rated at least three restaurants on Yelp. We verified that the participants indeed reviewed at least three restaurants using the participants' Yelp profile (Sharpe Wessling, Huber, and Netzer 2017). Having access to the reviewer's Yelp profile page also enabled us to collect the exact number of restaurant reviews each reviewer had written in the past, how long they had been a Yelp member, and their review distribution. Using a short survey, we asked these reviewers how often they go to restaurants per month and how many restaurants they visited in the last month for the first time. We then divided the number of restaurant reviews each reviewer wrote per month by the number of (1) sit-down restaurants that the reviewer visits per month and (2) sit-down restaurants visited for the first time in the last month. These ratios give us measures of the proportion of restaurants a reviewer reviewed relative to the restaurants visited—direct measures of polarity self-selection.

Similar to the analysis conducted in Figure 3, for reviewers who completed our survey ($n_{reviewers} = 95$, $n_{reviews} = 3{,}233$),[17] we find that reviewers in the upper quartile of the ratio exhibit a nonpolar distribution of reviews, but reviewers in the lower quartile display a polar distribution (right two histograms in Figure 3, Panel B). Comparing the histograms in both panels of Figure 3, we find that using either the ratios of reviews or number of reviews leads to strikingly similar patterns, suggesting that the number of reviews a reviewer wrote is a good proxy for polarity self-selection. Admittedly, the number of reviews a reviewer writes, or even the proportion of reviews to restaurants visited, is a proxy for the broader concept of self-selection, which includes polarity self-selection but may also include intertemporal self-selection or purchase self-selection. However, as the previous analyses show, this proxy of self-selection is indeed related to the polarity of the review distribution, which is consistent with polarity self-selection.

In addition to validating the number of reviews per reviewer as a proxy for polarity self-selection, we use the survey-based measure for polarity self-selection to contrast our proxy for polarity self-selection with two alternative proxies that can be obtained from secondary data: the median time between reviews as well as the variance of the interreview times. It can be assumed that when the interreview time is longer or when there is high variation in interreview time, self-selection is higher. We regress the proportion of the number of reviews written per month to the number of restaurants visited per month on the three proxies for polarity self-selection, independently and together (see Table 5). We find that the number of reviews per reviewer explains as much as 74% of the variation in our survey measure of self-selection (Model 1). The two interreview time measures explain only 31% of the variation (Models 2 and 3). Putting all three measures together, we find that only the number of reviews per reviewer is significant (Model 4). An incremental F-test shows that neither the median days between reviews ($F(1, 92) = .5192$, $p = .473$) nor the standard deviation of days between reviews ($F(1, 92) = .0102$, $p = .920$) add extra explanatory power over and beyond the number of reviews per reviewer. Thus, we conclude that the number of reviews per reviewer on its own is a good proxy for self-selection.

To further examine how well the number of reviews per reviewer measured from the secondary data as a proxy of self-selection relates to direct elicitation of self-selection measured by the survey responses, we regressed polarity and positive imbalance on the number of reviews per month (proportion of number of reviews divided by the membership duration in months; Model 1), the proportion of the number of

[17] We removed 33 participants because they either (1) had fewer than three reviews on Yelp or (2) created an account and wrote three reviews on the day of the survey only to participate in the survey.

Table 5. Analysis of Polarity Self-Selection Based on the Yelp Reviewers Survey.

DV: Survey-Based Self-Selection (Proportion of Restaurants Reviewed)				
	Model 1	Model 2	Model 3	Model 4
Polarity self-selection Proxy: log(reviews per month)	1.047 (.064)***			1.011 (.094)***
Polarity self-selection Proxy: median days between reviews		−.005 (.001)***		−.446 (.628)
Polarity self-selection Proxy: standard deviation days between reviews			−.003 (.000)***	.005 (.321)
Intercept	−1.340 (.130)***	−2.471 (.154)***	−2.185 (.183)***	−1.352 (.133)***
R^2	.742	.314	.309	.743
N	95	95	95	95

***$p < .01$.
Notes: SEs in parentheses.

Table 6. Analysis of Polarity Self-Selection Based on the Yelp Reviewers Survey.

		Model 1	Model 2	Model 3	Model 4	Model 5
Polarity self-selection Proxy: log(reviews per month)	Polarity	−.059 (.015)***				
	Pos. imbalance	−.055 (.014)***				
Polarity self-selection Proxy: log(ratio of number of reviews to restaurants visited per month)	Polarity		−.039 (.012)***			
	Pos. imbalance		−.043 (.011)***			
Polarity self-selection[a] Proxy: log(ratio of number of reviews to new restaurants visited per month)	Polarity			−.038 (.014)***		
	Pos. imbalance			−.040 (.013)***		
Log(number of restaurants per month)	Polarity				.013 (.021)	
	Pos. imbalance				.005 (.003)*	
Log(number of new restaurants per month)[a]	Polarity					−.005 (.026)
	Pos. imbalance					−.001 (.008)
Intercept	Polarity	.290 (.033)***	.273 (.043)***	.306 (.040)***	.383 (.035)***	.406 (.025)***
	Pos. imbalance	.343 (.032)***	.305 (.039)***	.345 (.038)***	.416 (.023)***	.451 (.026)***
Number of friends (in 1,000s)	Polarity	.021 (.206)	.001 (.211)	−.076 (.217)	.063 (.223)	.064 (.230)
	Pos. imbalance	.112 (.196)	.081 (.196)	.006 (.203)	.159 (.206)	0.150 (.216)
Number of followers (in 1,000s)	Polarity	.376 (4.906)	.043 (5.014)	−.367 (5.116)	−2.254 (5.248)	−3.026 (5.269)
	Pos. imbalance	.428 (4.670)	.560 (4.652)	.045 (4.793)	−1.937 (4.880)	−2.728 (4.983)
Number of years reviewer received Elite badge	Polarity	−.002 (.013)	−.009 (.013)	−.005 (.014)	−.024 (.013)*	−.021 (.013)
	Pos. imbalance	−.004 (.012)	−.008 (.012)	−.005 (.013)	−.023 (.012) *	−.022 (.012)*
R^2 polarity		.218	.184	.152	.088	.077
R^2 positive imbalance		.216	.223	.174	.125	.080
N		95	95	89[a]	95	89[a]

*$p < .1$.
**$p < .05$.
***$p < .01$.
[a]In this analysis, we exclude six respondents who indicated that they had not visited any restaurants for the first time in the last month.
Notes: SEs in parentheses.

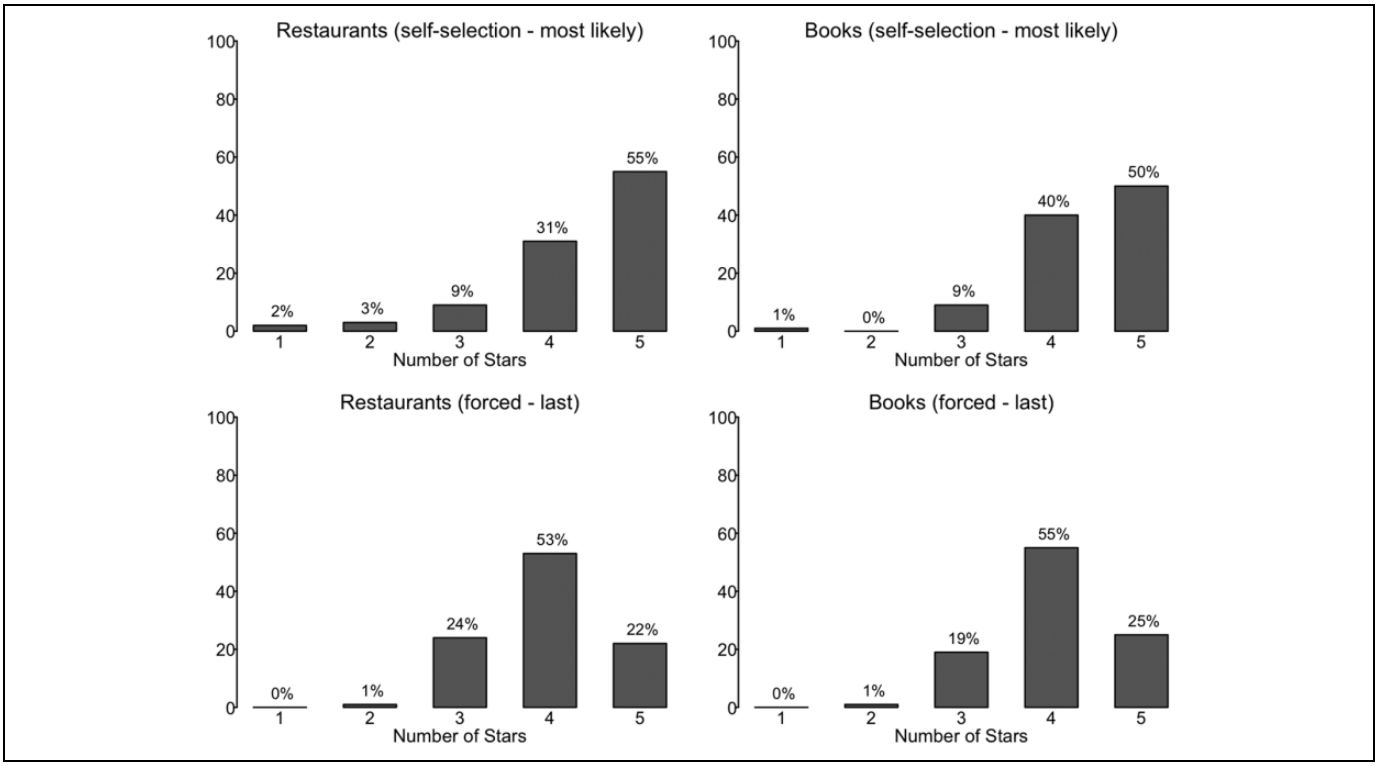

Figure 4. Empirical distributions for "self-selected" versus "forced" reviews.

reviews written per month to the number of restaurants visited per month (Model 2), the proportion of the number of reviews written per month to the number of restaurants visited per month for the first time (Model 3), the number of restaurants visited per month (Model 4), and the number of restaurants visited for the first time per month (Model 5).

As we expected, consistent with the cross-platform analysis we find that a high number of reviews per reviewer is correlated with higher polarity and positive imbalance (see Model 1 in Table 6). Measuring self-selection directly by replacing the frequency of reviews with the ratio of reviews written about restaurants visited (inverse polarity self-selection) leads to similar results (Models 2 and 3). In addition, our results suggest that an alternative explanation that polarity and positive imbalance of the review distribution is driven by different reference points of individuals who are frequent restaurant visitors versus individuals who rarely go to restaurants does not seem to explain the polarity and positive imbalance of reviews, as the relationship between restaurant visits and polarity and positive imbalance are insignificant or only marginally significant (Models 4 and 5).

We also tested alternative versions of Model 1 with the two additional proxy for self-selection (median interreview time and the variance of the interreview time). We find that while these measures are significantly related to polarity and positive imbalance both using the survey data and using the larger secondary Yelp data, the number of reviews per reviewer is a stronger predictor of both polarity and positive imbalance (for details, see Web Appendices 7 and 8). Taken together, these results point to the robustness of the polarity self-selection effect using three alternative measures. In addition, we find that the number of reviews per reviewer is a good proxy for polarity self-selection and is superior to alternative measures.

Manipulating Polarity Self-Selection via Experimental Design

Our analysis of secondary data from millions of reviews, complemented by survey data of Yelp reviews, provides strong evidence for polarity self-selection. We further complement our previous analyses with two (restaurant and book reviews) between-subjects-design experiments in which we manipulate polarity self-selection in a controlled environment. We use two between-subjects conditions: forced condition (last restaurant visited [book read]) and polarity self-selection condition (restaurant [book] most likely to review). $N_{restaurants} = 149$ (61% female) and $N_{books} = 158$ (56% female) Master's students from a large European university participated in these experiments for a course credit. Participants were randomly assigned to the polarity self-selection condition or the forced condition. In the forced condition, we ask participants to write a review about the last restaurant they have visited (book they have read). This condition should provide a review of a "randomly" selected product, which happens to be the last product participants experienced. Thus, this condition should be free of polarity self-selection because it does not permit the reviewer to select which product to review. In the polarity self-selection condition, we aim to mimic the typical online review environment

Table 7. Relationship Between Polarity and Time Since Purchase.

DV: Polarity = 1, No Polarity = 0	Restaurants (N = 149)	Books (N = 158)
Intercept	−2.645 (.566)***	−1.552 (.417)***
Manipulation (1 = polarity self-selection condition)	1.113 (.390)***	1.020 (.356)***
Time since purchase	.564 (.196)***	.164 (.116)

***$p < .01$.
Notes: SEs in parentheses.

and ask participants to write a review for a restaurant (book) for which they have written a review for in the past, or, if they have never written a review for a restaurant (book), for the restaurant (book) they would be most likely to write a review for. We chose restaurants and books as the two product categories because (1) they can highlight potential differences between services and products, (2) reviews for books and restaurants have been commonly used in previous research, and (3) consumer interest in restaurant and book reviews is overall high. We use the typical Amazon five-point-scale rating format (for the experimental stimuli and randomization check, see Web Appendix 9).

The review distributions in Figure 4 reveal that the polarity self-selection ("most likely") condition leads to a polar distribution of reviews, while the forced ("last") restaurant (book) condition leads to a distribution with a mass at the fourth scale point. The differences between the two distributions are statistically significant when comparing polarity (the proportion of one- and five-star reviews; $\chi^2_{\text{restaurants}}(1, N = 149) = 18.014$, $p < .001$; $\chi^2_{\text{books}}(1, N = 158) = 11.582$, $p < .001$), as well as the overall distributions (Fisher exact test [two sided]: restaurants: $p < .001$; books: $p = .004$). Thus, forcing individuals to write a review about their last experience and thereby eliminating polarity self-selection creates a nonpolar distribution of consumer reviews (bottom graphs of Figure 4), while allowing individuals to self-select the reviewed product creates a polar distribution (top graphs of Figure 4). We do not find a significant difference between the two conditions for positive imbalance ($\chi^2_{\text{restaurants}}(1, N = 149) = 1.287$, $p = .257$; $\chi^2_{\text{books}}(1, N = 158) = .004$, $p = .950$).[18]

One possible confound with the design of our experiment is that the two conditions might imply a different time frame. Whereas the reviews in the "forced" condition are written for a recent experience, the self-selected reviews can refer to an experience that occurred a long time ago. This might lead to differences in the reported review ratings. In the experiment, we asked respondents in both conditions how long ago they visited the restaurant (read the book) they reviewed. To investigate the potential impact of the time since purchase, we regress the measure of polarity on the experimental condition (coded as 1 for the polarity self-selection condition and 0 for the forced condition) controlling for the time since the product was purchased. The results indicate that after controlling for the time since purchase, polarity is significantly higher in the polarity self-selection condition (see Table 7). Moreover, we find no significant effect of the time since purchase for books. For restaurants, we find that the longer ago the experience is in the past, the higher the polarity of reviews. Thus, if anything, the time since purchase enhances the polarity of the distribution. To further examine the effect of review timing, we ran a separate study in which we included only the "forced" condition but split respondents on the basis of their reported likelihood to actually review that book/restaurant. That study holds time of experience constant. We find similar results to the one reported previously (i.e., a more polar distribution of reviews for books/restaurants that were more likely to be reviewed relative to those that were less likely; see Web Appendix 11).

In summary, we corroborate that, in a controlled experimental setting, polarity self-selection influences the polarity commonly observed in online review distributions. We note that although purchase self-selection may cause a customer to both buy the product and be more likely to rate it, relative to a customer who did not buy the product, purchase self-selection cannot lead to the difference we observe in our experiment because the respondents "purchased" the product in both conditions. Thus, we hold purchase self-selection constant across conditions.

The Role of Social Influence

Another possible force that may affect the shape of review distributions is social influence (Bikhchandani, Hirshleifer, and Welch 1992). Previous research has documented that the decision to review and the evaluation provided can change over time due to social influence factors such as previous reviews (e.g., Godes and Silva 2012; Moe and Schweidel 2012; Moe and Trusov 2011; Muchnik, Aral, and Taylor 2013; Sridhar and Srinivasan 2012) or the composition of reviewer groups (Li and Hitt 2008; Moe and Schweidel 2012). Social influence in the form of previous reviews can also influence the consumer's expectations of product value (Li and Hitt 2008). In addition, existing ratings can influence one's motivation to review and bias the ratings consumers provide in the postpurchase stage (Moe and Schweidel 2012). Therefore, social influences can enhance polarity self-selection via two routes at different points in the purchase and review funnel: (1) before buying the product (i.e., social influences can create [high] expectations of consumers, leading to a higher polarity self-selection to rate) and (2) before rating the product (i.e., the existing ratings of other consumers can influence both the incidence decision to review a product and the product evaluation the consumer provides, thereby biasing the rating).

Unlike studies that have investigated social influence via experimental settings (Schlosser 2005) or secondary data (Moe and Schweidel 2012), our polarity self-selection experiments did not expose participants to any previous ratings prior to

[18] We replicated this study with MTurk participants (N = 100) adding an additional category (movies) and obtained very similar results (see Web Appendix 10).

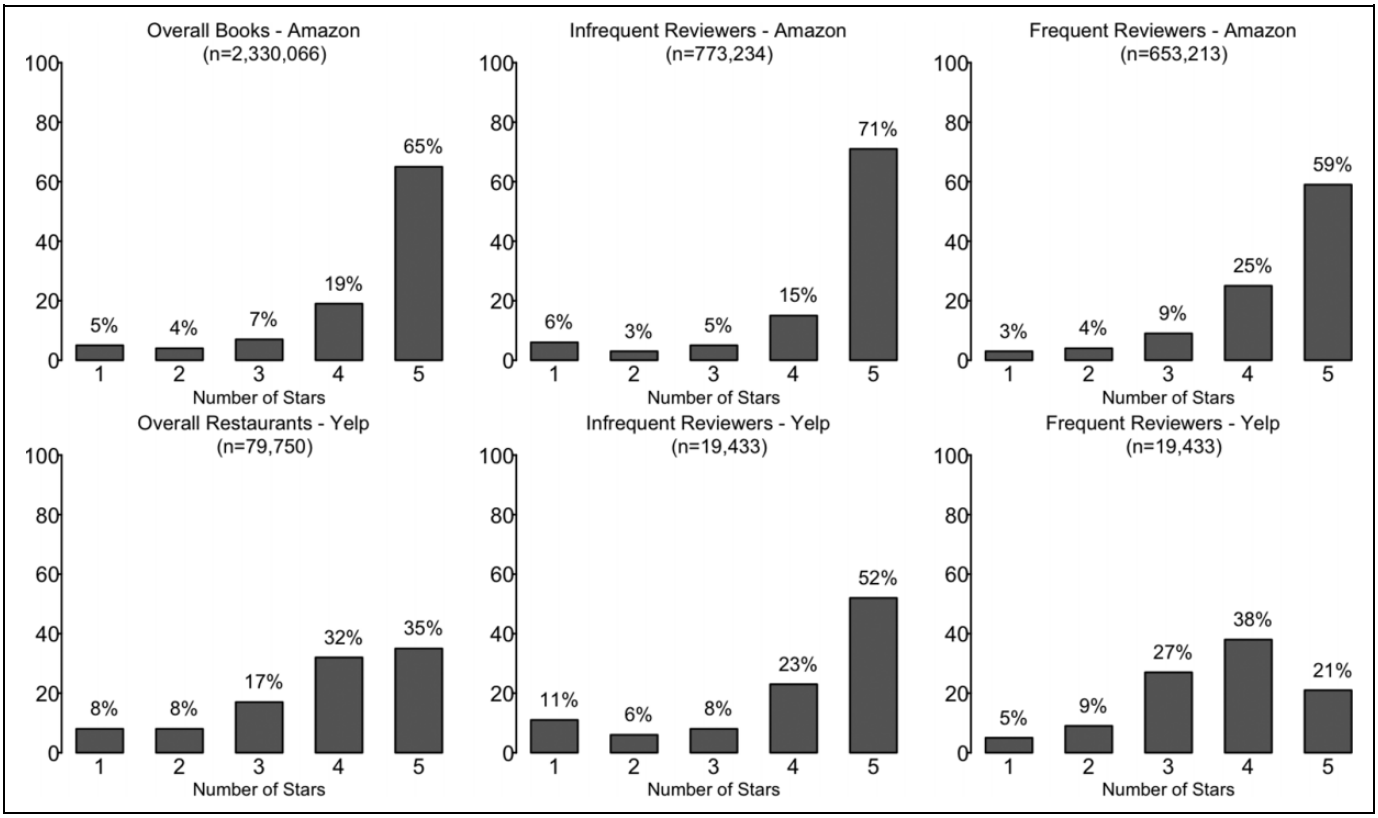

Figure 5. Review distribution for first review per book or restaurant.

Table 8. Logistic Regression for Polarity in the First Product Review and Number of Reviews per Reviewer.

	Yelp[a] (N = 79,750)	Amazon (N = 2,330,066)
Log(number reviews of reviewer)	−.375 (.005)***	−.112 (.001)***
Intercept	−.443 (.008)***	1.146 (.002)***

***$p < .01$.

[a]For Yelp, we use the number of restaurant reviews per month because we have the time the user joined the platform. We find similar results if we use the overall number of reviews. Because we do not have this information for Amazon, we use the reviewer's number of reviews.

Notes: SEs in parentheses.

providing their ratings. Thus, the polarity self-selection we find in our experiments is unlikely to be affected by the existing rating environments.[19]

To further empirically investigate whether postpurchase social influences can contribute to the polarity and positive imbalance of the review distribution, we examine whether polarity and positive imbalance exist prior to any such social influence or dynamics. To do so, we look at the very first review across 2,330,066 Amazon book reviews and 79,750 Yelp restaurant reviews (see Figure 5). We see that even for the first book/restaurant review (prior to any temporal change) the distribution of reviews is both polar and positively imbalanced (see leftmost charts in Figure 5). This result is consistent with the findings of Li and Hitt (2008), who suggest that early adopters (and reviewers) of products are likely to be strong proponents of the product or experts in the category and thus are likely to rate products at the extreme. Consequently, even in the absence of any prior reviews—and thus the absence of social influence factors—we find a polar distribution.

To investigate whether polarity self-selection is present at the time of the first review, we compare the first review of products by splitting the first reviews into two groups according to the upper and lower quartiles of the number of reviews per reviewer who wrote the review. The middle and right charts in Figure 5 depict the quartile split. It is evident that first reviews written by infrequent reviewers exhibit more polar distributions relative to first reviews written by frequent reviewers.

To investigate the relationship between polarity self-selection and the first review distribution statistically, we use a continuous analysis. We ran a logistic regression of whether the review was polar (one- or five-star) or not on the number of reviews written by the reviewer. The results in Table 8 show an

[19] Admittedly, participants in our experiment may take into account the reviews they have seen in online platforms for the restaurant (book) they have visited (read). However, such a process would require respondents first to be exposed to the previous reviews and then to accurately recall these reviews at the time of the experiment.

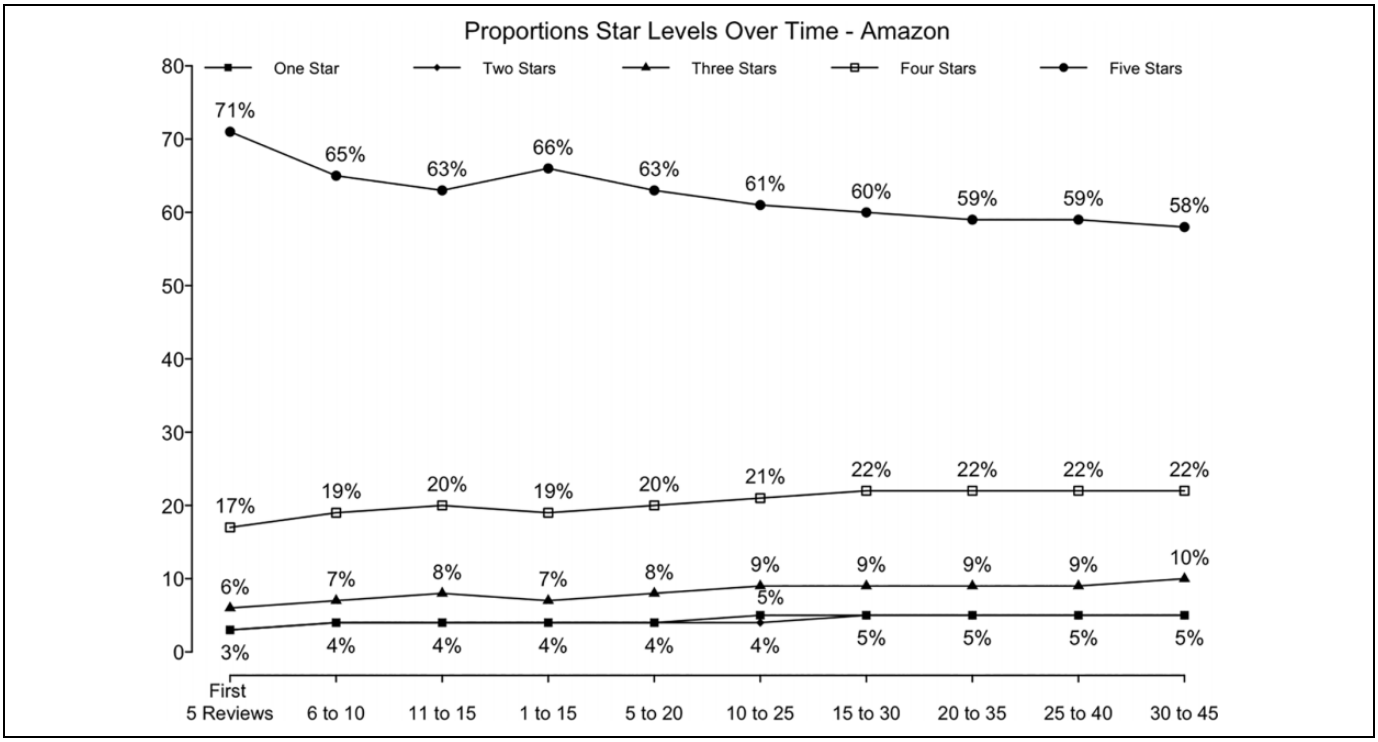

Figure 6. Review distribution in Amazon over subsequent reviews.

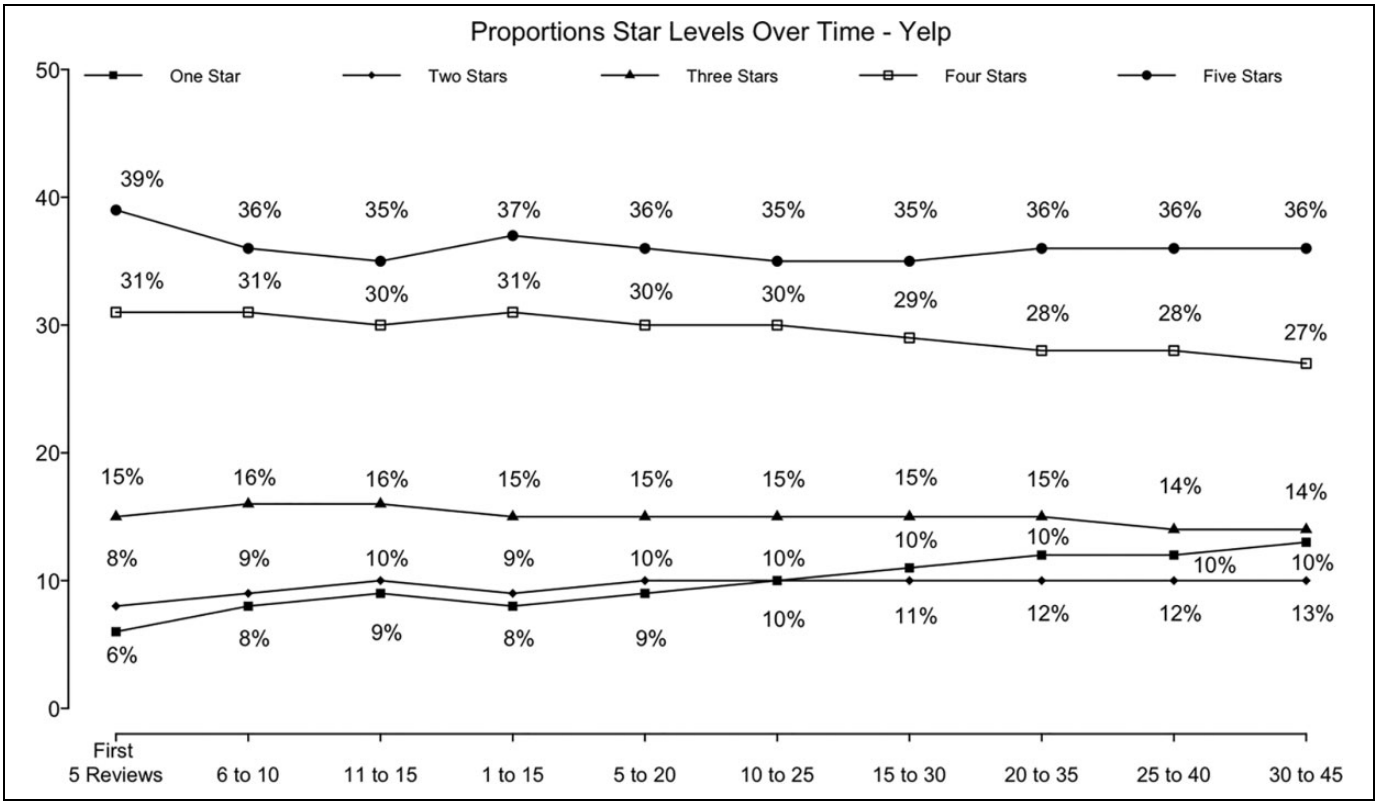

Figure 7. Review distribution in Yelp over subsequent reviews.

increasing number of reviews per reviewer leads to decreased polarity. These differences between frequent and infrequent reviewers are consistent with the polarity self-selection account and cannot be explained by social influence, as these were the first reviews ever written for the product.

To investigate whether social influence affects the review distribution for subsequent reviews following the initial review, we analyze the evolution of the review distribution over time. Specifically, we examine 79,535 Amazon books as well as 22,231 Yelp restaurants with at least 45 reviews each. To analyze systematic changes of the review distributions, we calculate the average proportion of review distributions across books (restaurants) using a moving window of 15 reviews increasing by an increment of 5 reviews at a time, up to 45 reviews. Specifically, we calculate the first distribution using the first 15 reviews, the second using the 6th to 20th reviews, and so on, with the last 15 reviews window including reviews ordered 31–45.

The reviews at all review order windows of Amazon books exhibit a polar distribution (the majority of the reviews are at the poles [one- and five-star ratings] throughout the time window; see Figure 6). For Yelp, we find a similar pattern—a decline of five-star ratings and an increase of one-star ratings (see Figure 7). The decline in five-star ratings over time due to social influence (Godes and Silva 2012; Li and Hitt 2008; Moe and Schweidel 2012; Moe and Trusov 2011) suggests that, if anything, social influence seems to reduce the polarity and thus cannot be the underlying driver.

Overall, we demonstrate that (1) polarity self-selection is present already in the first review, which, by definition, cannot be affected by social influence; (2) even during the first review, reviews are more polar for reviewers higher on polarity self-selection (fewer reviews per reviewer); (3) platforms seem to exhibit dynamics in the distribution of reviews over time, but if anything, these dynamics decrease the polarity of the distribution over time; and (4) we observe that the review distributions of both Amazon and Yelp exhibit polarity over the life cycle of the product.

Consequences of Polarity Self-Selection: Why the Average Ratings Metric May Be Misleading

Given the polarity of the distribution of online reviews, summarizing reviews with average ratings can mask important information. Nevertheless, despite the prevalence of polar review distributions in platforms such as Amazon, many platforms (and academic papers) use average ratings as a metric to inform consumers and relate online reviews to sales (e.g., Chevalier and Mayzlin 2006; Chintagunta, Gopinath, and Venkataraman 2010; Ghose and Ipeirotis 2011). Moreover, De Langhe, Fernbach, and Lichtenstein (2015) demonstrate that consumers tend to rely more heavily on average ratings than on the number of ratings in forming their quality evaluations.

Table 9. Descriptive Statistics Song Rating Data.

	Amazon	Survey
Mean number of ratings	50.21	37.33
Mean average rating	4.72	3.23
Polarity of ratings	87.34%	28.78%
Positive imbalance of ratings	95.96%	58.08%

Indeed, You, Vadakkepatt, and Joshi (2015, p. 20) state that the "failure to consider distribution intensity significantly affect[s] eWOM valence elasticities but not volume elasticities." This result is consistent with polarity self-selection. If mainly consumers who are satisfied with the product are likely to review it, as suggested by polarity self-selection, then the average rating is likely to be uninformative because most consumers who review the product rate it positively, making products indistinguishable based on average ratings. However, the number of reviews is likely to be informative even in the presence of high polarity self-selection, because if many consumers review the product it implies that many consumers were satisfied with it.

In this section, we therefore investigate the role of polarity self-selection in the lack of informativeness of the average rating metric in informing consumers in their product selection and purchase decisions. Specifically, we investigate the (lack of) predictive value of average reviews with respect to revealed preferences as represented by sales and objective product quality. If polarity self-selection reduces the information provided by average ratings, replacing the self-selected set of reviews with reviews that suffer from lower polarity self-selection should lead to a more accurate measure of average ratings and thus a stronger relationship between average ratings and sales as well as objective quality.

Reducing the Self-Selection in Calculating Average Ratings

In the following analysis, we examine whether the review valence measure can be enhanced by including reviews from a population that was required (as opposed to self-selected) to review the product. We investigate this in context of song reviews. We chose songs as a product category in this analysis because consumers are likely to be familiar with a large number of songs and thus able to rate many songs. We chose the top 50 songs from the category "most popular titles" on Google Play.[20] For the set of the 50 songs, we collected the online rating information on Amazon, the sales rank of the song on Amazon "paid songs" as well as the "publication" date of the song on Amazon. To allow for a reliable analysis, we eliminated 11 songs that had fewer than three ratings or were not accessible on Amazon. To obtain a measure of song ratings that

[20] We restrict our analysis to the top 50 songs that had ratings for the specific song, as opposed to overall album ratings.

Table 10. Relationship Between Log(Song Sales Rank) and Review Volume and Valence.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Log(Amazon # reviews)	−.666 (.158)***			−.674 (.161)***	−.573 (.162)***	−.581 (.163)***	
Amazon avg. reviews		−.047 (.957)		.339 (.798)		.580 (.780)	
Survey avg. reviews			−2.096 (.777)**		−1.299 (.712)*	−1.387 (.726)*	−1.575 (.799)*
Log(Google # reviews)							−.472 (.193)**
Google avg. reviews							−.438 (1.407)
Release date	.001 (.003)	−.001 (.003)	−.003 (.003)	.001 (.003)	−.001 (.003)	−.001 (.003)	−.007 (.003)**
Intercept	−20.701 (115.597)	57.141 (140.044)	124.968 (129.547)	−16.854 (117.286)	31.890 (115.700)	42.010 (117.200)	311.880 (142.656)**
N	39	39	39	39	39	39	39
R^2	.333	.004	.172	.336	.391	.401	.304

*$p < .1$.
**$p < .05$.
***$p < .01$.
Notes: SEs in parentheses.

is less affected by polarity self-selection, we collected a survey of MTurk workers (n = 146, age range 18–35 years) who were asked to rate a set of 25 songs (from one of two randomly chosen sets of the total of 50 songs). For songs that respondents were not familiar with, we added the choice alternative "don't know the song." On average, participants rated 51.27% of the songs included in the sample of the 39 songs as "don't know the song," leading to 1,456 song ratings (an average of 37.33 ratings per song). We then compare the survey-based ratings with the ratings on Amazon. To mitigate the concern that, at the time of ratings, survey respondents were already aware of the song's popularity, we include in our analysis only Amazon reviews that were posted after the time of the survey.[21] Table 9 shows the descriptive statistics of the Amazon ratings and the survey ratings. Consistent with polarity self-selection, the survey average ratings are significantly lower relative to Amazon's ratings and exhibit a considerably lower polarity and positive imbalance.

We first assess the relationship between Amazon's sales rank and the number and average rating of Amazon reviews as well as the survey ratings. We regress the log of the sales rank of the song on Amazon on each variable separately controlling for the Amazon release date (see Models 1, 2 and 3 in Table 10). For the interpretation of the results, note that songs with higher sales have a lower sales rank. Interestingly, despite the fact that consumers buying songs on Amazon were most likely exposed to the Amazon reviews and the fact that our survey results were based on a nonrepresentative MTurk sample, which may not be representative of the Amazon Music customer base, we find that the average rating of the survey data explains much more of the variation in the Amazon sales rank than the average Amazon reviews ($R^2_{\text{survey avg}} = .172$ vs. $R^2_{\text{Amazon avg}} = .004$).

Including the variables together in the same regression (Model 4), we find that, consistent with previous studies (Forman, Ghose, and Wiesenfeld 2008; Hu, Koh, and Reddy 2014), when we include both the number of ratings and the average ratings on Amazon, only the number of reviews is significantly related to the sales rank. This result is also consistent with the strong polarity self-selection on Amazon. Next, we replace the average rating of Amazon reviews with the average rating from our survey (Model 5). While the log number of reviews is still significant in this model, the average rating from the survey is also (marginally) significantly related to the sales rank. In addition, the explained variation in sales rank is considerably higher when we replace the average rating on Amazon with the average rating from the survey ($R^2_{\text{survey avg}} = .391$ vs. $R^2_{\text{Amazon avg}} = .336$). This result holds also when we include all three predictors in the model (Model 6). To examine if the problem lies with the Amazon ratings, we also used Google Play's average ratings and number of reviews for the respective songs instead of Amazon's ratings and number of reviews (Model 7). Only the number of reviews on Google Play and the average

[21] Analysis with all Amazon reviews led to similar results.

Table 11. Relationship Between Critic Reviews and Polarity Self-Selection.

Intercept	105.846 (67.925)
Log(average # of reviews per reviewer)	−35.054 (12.100)***
IMDb average rating	−5.317 (10.039)
Log(average # of reviews per reviewer) × average ratings	5.050 (1.796)***
N	150
R^2	.650

***$p < .01$.
Notes: SEs in parentheses.

Table 12. Relationship Between Review Usefulness and Polarity Self-Selection.

Intercept	.976 (.011)***
Log(number restaurant reviews by reviewer per month)	.200 (.001)***
Log(number restaurant reviews)	.022 (.001)***
Average ratings	.045 (.003)***
N	4,467,240
R^2	.015

***$p < .01$.
Notes: SEs in parentheses.

rating from the survey are (marginally) significantly related to song sales. This again confirms that polarity self-selection leads to a loss of information of the average rating.

The Relationship Between Self-Selection and Objective Ratings as Well as Usefulness

Arguably, online reviews should be most closely related to objective quality of the product such as those reflected in product evaluations provided by experts (e.g., consumer reports ratings, expert movie critics). However, previous research has shown only a weak relationship between average ratings and objective product quality measures for Amazon online reviews (De Langhe, Fernbach, and Lichtenstein 2015). Could the low relationship found in De Langhe, Fernbach, and Lichtenstein (2015) be attributed to the high polarity self-selection on the Amazon platform? Next, we evaluate the role of polarity self-selection in explaining this discrepancy. Specifically, consistent with polarity self-selection, we should expect a stronger relationship between average online review ratings and objective evaluation for products for which polarity self-selection is low relative to products for which polarity self-selection is high.

To examine this, we collect a data set of movies from IMDb and combine it with movie critic ratings from Rotten Tomatoes as a measure of objective quality. Although critic ratings offer a better objective measure of product quality compared with consumer ratings, they should not be seen as a ground truth measure of product quality, but rather a proxy for objective quality. We collect the top 150 movies released in the United States in 2017. As a measure of self-selection, we gathered the number of reviews of reviewers for a random sample of 100 reviewers (written by unique reviewers) per movie. We average the number reviews per reviewer per movie to assess the polarity self-selection of a movie.

We first correlate the Tomatometer score of movie critics from Rotten Tomatoes with the average online ratings from IMDb. We find that the correlation between the expert's Tomatometer score and IMDb consumer reviews ($r(150) = .7935$, $p < .001$) is much larger than the correlation between the expert's Tomatometer score and the same movie consumer reviews on Amazon ($r(150) = .3738$, $p < .001$). This result is anecdotally in line with the higher polarity self-selection on the Amazon platform (4 reviews per reviewer) relative to the IMDb platform (367 reviews per reviewer). This result may suggest that previous findings on the weak relationship between average ratings and expert ratings (De Langhe, Fernbach, and Lichtenstein 2015) may be attributed to the choice of a high-self-selection platform (Amazon).

To assess the relationship between self-selection and the relationship between average ratings and objective quality further, we regress the Tomatometer score of movie critics on the average rating, the number of reviews per reviewer, and the interaction between the two (see Table 11). We find a significant positive interaction effect. This result demonstrates that average ratings are more strongly related to critic ratings, and thus objective quality when self-selection is lower (number of reviews per reviewer is higher).

Another dimension of review informativeness is review's perceived usefulness, which relates to the perception of reviews by (potential) consumers. Previous research has shown that extreme reviews are often perceived as less useful (Mudambi and Schuff 2010). We build on these findings and investigate the impact of polarity self-selection on the perceived usefulness of Yelp reviews. Again, we measure polarity self-selection of a reviewer by the number of reviews they have written per month. To test the effect of polarity self-selection on usefulness of the review, we regress usefulness as measured by the number of usefulness votes of a review on Yelp on the number of reviews of a reviewer, controlling for the overall number of reviews of the restaurant and the average rating. The regression results are summarized in Table 12. We find that polarity self-selection has a significant negative effect on usefulness. That is, reviews written by reviewers who write more reviews (lower self-selection) are perceived as more useful.

Overall, our analyses demonstrate that polarity self-selection can distort the informativeness of average review ratings and, at the same time, increase the informativeness of the number of reviews as an indicator of product popularity and objective quality measures. We further demonstrate that complementing the existing (self-selected) reviews with a less self-selected source of evaluation in the form of survey-based reviews leads to a more reliable measure of consumers' preferences as reflected by products' sales. The role of polarity self-selection with respect to the polarity of the review distribution highlights the importance of understanding not only

average ratings but also the entire review distribution in investigating the relationship between reviews and sales.

Conclusions and Future Directions

Online reviews are a major source of information for consumers in making decisions. This is reflected by the sheer number of academic and nonacademic studies investigating the impact of online reviews on consumers' decision making. Prior research has documented a high degree of polarity and positive imbalance of review distributions. In this research, we investigate the prevalence of the degree of polarity and positive imbalance of online review distributions, the role of polarity self-selection in generating this distribution, and, finally, its consequences with respect to the loss of informativeness of the average review measure. We conduct a comprehensive analysis of the prevalence of the polarity of review distributions across more than 280 million reviews from 25 online review and e-commerce platforms. We find that while polarity in the review distribution is quite prevalent, substantial variation exists across platforms. Platforms in which reviewers only review a few products on average (high degree of polarity self-selection) exhibit high degrees of polarity relative to platforms in which reviewers review many products. We demonstrate that the number of reviews per reviewers can serve as a good, and easy to collect, proxy for polarity self-selection and propose it as a measure for platforms to identify products and/or reviewers with a high degree of polarity self-selection.

We use a multipronged approach including secondary data analyses of a large-scale review database, experiments, and surveys to investigate the role of polarity self-selection in explaining the degree of polarity of the review distributions. Using experiments, we demonstrate the causal relationship between polarity self-selection and the polarity of the distribution.

Finally, we find that polarity self-selection can lead to a loss of information of online ratings. We show that for review distributions with a higher intensity of polarity self-selection, the average ratings metric, commonly used in online review platforms and academic research, is only weakly related to sales. In addition, we suggest that the inconclusive results in previous research regarding the relationship between the average rating and sales can be explained by polarity self-selection and the prevalence of polar distributions of reviews. We further demonstrate that while the average ratings of the reviews on the Amazon platform are only weakly related to sales, replacing these reviews with a non-self-selected set of reviews leads to a stronger relationship between average ratings and sales.

Our results have important implications for both platforms that host online reviews and consumers that use these reviews as a source of information. We demonstrate that due to polarity self-selection, reviews posted online reflect a possibly biased view of the overall evaluations of consumers who purchased the product. Consumers using reviews as a source of online WOM should be aware that reviews on many platforms reflect an extreme picture of the true shape of consumer preferences. At the extreme, when most reviews are polar and positively imbalanced, consumers should pay more attention to the number of reviews than to the average rating, as the average ratings do not allow to distinguish between high- and low-quality products.

Platforms aiming to provide consumers with an unbiased view of consumers' evaluations and to increase the informativeness of their online reviews should strive to assess the degree of polarity self-selection, and, to the extent possible, reduce it and inform consumers about its existence. We propose the number of reviews per reviewer as not only a good proxy for self-selection, but also an easy measure for platforms to collect and approximate the degree of self-selection of the platform's reviewers. One potential strategy for platforms to increase the informativeness of reviews might be to report the average or the distribution of the number of reviews of reviewers that rated the product. For example, platforms could allow consumers to compare review statistics and distributions for frequent versus infrequent reviewers. One online platform that does not present the raw average ratings but weighs the ratings prior to showing the average is IMDb. Although IMDb does not disclose the weighing algorithm (IMDb 2020), our research suggests that in such an algorithm the reviews of frequent reviewers should receive a higher weight than those of infrequent reviewers.

Another strategy for platforms to reduce polarity self-selection is to increase customers' propensity to write a review. Platforms can send out reminders or even monetarily incentivize customers to write reviews about products they experience. To investigate how incentivizing reviewers may affect the distribution of reviews, we collaborated with the German service platform yourXpert and ran a field experiment incentivizing customers to write reviews. To test whether review platforms sending reminders to review a product or service can reduce polarity self-selection, we use the following setup: the company sent a reminder to write a review two days after the purchase of the service, and after two weeks a second reminder to customers who did not rate the service after the first reminder with an offer of €5 if they review the service. We compare the share of polar ratings (one- and five-star ratings) for reviews that were received before the first reminder ($M = 90.4\%$; $N = 1{,}343$), reviews received after the first reminder without monetary incentive ($M = 83.5\%$; $N = 782$), and reviews received after the second reminder with a monetary incentive ($M = 79.9\%$; $N = 645$). We find that the proportion of one- and five-star ratings is significantly lower for reviews that arrived after the first reminder without monetary incentive ($\chi^2(1, N = 2{,}125) = 21.358$, $p < .001$) and for reviews that arrived after the second reminder with monetary incentive ($\chi^2(1, N = 1{,}988) = 41.873$, $p < .001$) relative to reviews that arrived with no reminder. These results provide initial evidence that incentivizing customers to write reviews for products they have purchased (with or without monetary incentive) can reduce polarity self-selection. We encourage future research to explore the role of monetary and other incentives in reducing

polarity self-selection and provide a more representative set of reviews (e.g., Brandes, Godes, and Mayzlin 2019).

Our findings regarding the role of polarity self-selection in the generation of reviews can help explain several findings reported in the online review literature. First, our results are consistent with the finding of Moe and Schweidel (2012), who demonstrate that frequent reviewers are, on average, more negative than infrequent reviewers. The authors explain this finding through the frequent reviewers' objective to differentiate themselves from the overall positive rating environment, whereas infrequent reviewers show bandwagon behavior, thus giving more positive ratings. Our finding suggests that the difference between the behavior of frequent and infrequent reviewers already exists in the first review they write for a product and may be related to the degree of polarity self-selection of these two groups of reviewers. Mayzlin, Dover, and Chevalier (2014) find that approximately 23% of all TripAdvisor reviews are posted by one-time reviewers and, consistent with our findings, show that these reviews are more likely to be polar compared with the entire TripAdvisor sample. While the authors explain this result by stating the possibility that one-time reviewers are more likely to write fake reviews, we show that this effect can be attributed to polarity self-selection.

We note several possible limitations of our research. First, although we investigate a large set of factors that can affect the variation in the polarity and positive imbalance of review distributions across platforms, other cross-platform factors, which we could not access at scale in the current analysis, might also affect the review distributions. We encourage future research to investigate the impact of factors such as whether and how much the reviewed product/service also advertises on the platforms, information from the reviewer's previous reviews on the platform, the strength and structure of the social network among reviewers, and the textual content of the reviews.

Second, we find evidence that the key review metrics (average ratings and number of reviews) may be biased by polarity self-selection and that such a bias may hinder the ability of these metrics to capture consumers' "true" preferences. Obtaining access to individual-level data of both reviews and purchases may help shed more light on the relationship between polarity self-selection and consumer choices. Future research could investigate the impact of review distributions on individual purchasing behavior.

Third, we document the prevalence of polarity and positive imbalance and identify polarity self-selection as a major factor in the observed distribution of reviews. However, in the scope of this article, we do not directly explore *why* consumers are more likely to review extreme experiences (Gershoff, Mukherjee, and Mukhopadhyay 2003). Future research should further explore the motivation to write reviews and its relationship to polarity self-selection. Polarity self-selection could arise from (1) prepurchase expectations (consumers self-select whether to review a product based on their expectation prepurchase), (2) the actual evaluation (consumers actual evaluation of the product after purchasing/experiencing it affects their decision to review), or (3) expectation disconfirmation—the difference between 1 and 2 (i.e., the change between expectations and the final evaluation; Minnema et al. 2016). The distinction between the three drivers can inform whether self-selection takes place before or after the customer has purchased the product. While expectations can be formed by existing product reviews, offline WOM, or advertising, expectation disconfirmation can be formed by the experience with the product and biases such as cognitive dissonance.

We ran a preliminary study in which we asked respondents about their expectation, expectation disconfirmation, and actual evaluation of a restaurant or a book as well as their likelihood to review. We find that all three drivers—prepurchase expectations, expectation disconfirmation, and overall ratings of the product—affect the customer's decision to write a review. The finding that prepurchase expectations affect the decision to review enables us to rule out the possibility that polarity self-selection is purely driven by adjustment-type biases, such as cognitive dissonance (Festinger 1957) or social influence. Interestingly, splitting the effect of expectation disconfirmation between positive and negative disconfirmation can help explain our positive imbalance results. We find that consumers with positive expectation disconfirmation were more likely to review compared with those with no or negative confirmation. From a managerial perspective, these results also suggest that companies setting expectations too high do not get penalized by worse-than-expected experiences when it comes to online reviews. Rather, such companies simply end up with fewer reviews, as only satisfied consumers seem to write reviews. This points to the importance of the number of reviews as a measure of overall quality. Alternatively, companies could set expectations low and thus "encourage" consumers to write a review due to positive expectation disconfirmation.

Fourth, the present study focuses on numerical ratings. Consumer reviews generally consist of numerical ratings and textual evaluations. Prior research has shown the relevance of review text for sales predictions (Archak, Ghose, and Ipeirotis 2011). Our initial examination reveals that the sentiment of review text shows lower polarity than the numerical ratings. We encourage future research to investigate the underlying sentiment of review text and identify potential differences between the numerical rating distributions and the sentiment of textual information.

Fifth, possible variation in scale usage across cultures could lead to differences in the role of polarity self-selection and the resulting review distribution (Li et al. 2018). Because most of the platforms we study are U.S.-based, we cannot rigorously analyze cross-cultural effects. We leave this interesting avenue of exploration for future research.

In summary, our research provides one of the largest-scale analyses of online reviews to explore the distributions of online reviews, the prevalence of polarity in these distributions, and the possible antecedents and consequences of such polarity. We identify that the summaries of online reviews—valence and volume—that are commonly used by consumers may be biased

due to polarity self-selection. We hope to encourage future works to further explore this bias and its possible remedies.

Acknowledgments

The authors thank the Wharton Customer Analytics Initiative; Julian McAuley; the Stanford Network Analysis Project; Stefan Schütt and Bernhard Finkbeiner from yourXpert and Frag-Mutti.de; Yaniv Dover, Dina Mayzlin, and Judith Chevalier for their help with providing data for this article; and Stefan Kluge for help with data collection.

Associate Editor

Hari Sridhar

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Verena Schoenmueller gratefully acknowledges support from the Swiss National Science Foundation.

References

Anderson, Eric T. and Duncan I. Simester (2014), "Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception," *Journal of Marketing Research*, 51 (3), 249–69.

Anderson, Eugene W. (1998), "Customer Satisfaction and Word of Mouth," *Journal of Service Research*, 1 (1), 5–17.

Anderson, Eugene W. and Claes Fornell (2000), "Foundations of the American Customer Satisfaction Index," *Total Quality Management*, 11 (7), 869–82.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis (2011), "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science*, 57 (8), 1485–1509.

Babić Rosario, Ana, Francesca Sotgiu, Kristine De Valck, and Tammo H.A. Bijmolt (2016), "The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors," *Journal of Marketing Research*, 53 (3), 297–318.

Barasch, Alixandra and Jonah Berger (2014), "Broadcasting and Narrowcasting: How Audience Size Affects What People Share," *Journal of Marketing Research*, 51 (3), 286–99.

Berger, Jonah (2014), "Word of Mouth and Interpersonal Communication: A Review and Directions for Future Research," *Journal of Consumer Psychology*, 24 (4), 586–607.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch (1992), "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100 (5), 992–1026.

Brandes, Leif, David Godes, and Dina Mayzlin (2019), "What Drives Extremity Bias in Online Reviews? Theory and Experimental Evidence," working paper.

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

Chintagunta, Pradeep K., Shyam Gopinath, and Sriram Venkataraman (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), 944–57.

Dalvi, Nilesh N., Ravi Kumar, and Bo Pang (2013), "Para 'Normal' Activity: On the Distribution of Average Ratings," in proceedings of *The International Conference on Weblogs and Social Media 2013*, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6117/6349.

Danaher, Peter J. and Vanessa Haddrell (1996), "A Comparison of Question Scales Used for Measuring Customer Satisfaction," *International Journal of Service Industry Management*, 7 (4), 4–26.

De Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2015), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42 (6), 817–33.

Dellarocas, Chrysanthos (2003), "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science*, 49 (10), 1407–24.

Dellarocas, Chrysanthos, Guodong Gao, and Ritu Narayan (2010), "Are Consumers More Likely to Contribute Online Reviews for Hit or Niche Products?" *Journal of Management Information Systems*, 27 (2), 127–58.

Dellarocas, Chrysanthos and Charles A. Wood (2008), "The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias," *Management Science*, 54 (3), 460–76.

Dellarocas, Chrysanthos, Xiaoquan (Michael) Zhang, and Neveen F. Awad (2007), "Exploring the Value of Online Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive Marketing*, 21 (4), 23–45.

East, Robert, Kathy Hammond, and Malcolm Wright (2007), "The Relative Incidence of Positive and Negative Word of Mouth: A Multi-Category Study," *International Journal of Research in Marketing*, 24 (2), 175–84.

Engel, James F., Robert J. Kegerreis, and Roger D. Blackwell (1969), "Word-of-Mouth Communication by the Innovator," *Journal of Marketing*, 33 (3), 15–19.

Feng, Song, Longfei Xing, Anupam Gogar, and Yejin Choi (2012), "Distributional Footprints of Deceptive Online Reviews," in proceedings of *The International Conference on Weblogs and Social Media 2013*, https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/download/4675/4972.

Festinger, Leon (1957), *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fisher, Matthew, George E. Newman, and Ravi Dhar (2018), "Seeing Stars: How the Binary Bias Distorts the Interpretation of Customer Ratings," *Journal of Consumer Research*, 45 (3), 471–89.

Floyd, Kristopher, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling (2014), "How Online Reviews Affect Retail Sales: A Meta-Analysis," *Journal of Retailing*, 90 (2), 217–32.

Forman, Chris, Anindya Ghose, and Batia Wiesenfeld (2008), "Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information Systems Research*, 19 (3), 291–313.

Fritz, Ben (2016), "Hollywood Now Worries About Viewer Scores, Not Reviews," *The Wall Street Journal* (July 20), http://www.wsj.com/articles/hollywood-turns-spotlight-on-websites-that-aggregate-movie-reviews-1469038587.

Gershoff, Andrew D., Ashesh Mukherjee, and Anirban Mukhopadhyay (2003), "Consumer Acceptance of Online Agent Advice: Extremity and Positivity Effects," *Journal of Consumer Psychology*, 13 (1/2), 161–70.

Ghose, Anindya and Panagiotis G. Ipeirotis (2011), "Estimating the Helpfulness and Economic Impact of Online Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering*, 23 (10), 1498–1512.

Godes, David and José C. Silva (2012), "Sequential and Temporal Dynamics of Online Opinion," *Marketing Science*, 31 (3), 448–73.

Heskett, James L., W.Earl Sasser Jr., and Leonard A. Schlesinger (1997), *The Service Profit Chain: How Leading Companies Link Profit and Growth to Loyalty, Satisfaction, and Value*. New York: The Free Press.

Hickey, Walt (2015), "Be Suspicious of Online Movie Ratings, Especially Fandango's," FiveThirtyEight (October 15), http://fivethirtyeight.com/features/fandango-movies-ratings/.

Hu, Nan, Jie Zhang, and Paul A. Pavlou (2009), "Overcoming the J-Shaped Distribution of Online Reviews," *Communications of the ACM*, 52 (10), 144–47.

Hu, Nan, Noi Sian Koh, and Srinivas K. Reddy (2014), "Ratings Lead You to the Product, Reviews Help You Clinch It? The Mediating Role of Online Review Sentiments on Product Sales," *Decision Support Systems*, 57, 42–53.

Hu, Nan, Paul A. Pavlou, and Jie Zhang (2017), "On Self-Selection Biases in Online Product Reviews" *MIS Quarterly*, 41 (2), 449–71.

Igniyte (2019), "30 Essential Online Review Facts and Stats," (October 29), https://www.igniyte.co.uk/blog/30-online-review-facts-and-stats/.

IMDb (2020), "Weighted Average Ratings," (accessed July 14, 2020), https://help.imdb.com/article/imdb/track-movies-tv/weighted-average-ratings/GWT2DSBYVT2F25SK#

Kaemingk, Diana (2019), "20 Online Review Stats to Know in 2019," Qualtrics (April 9), https://www.qualtrics.com/blog/online-review-stats/.

King, Robert A., Pradeep Racherla, and Victoria D. Bush (2014), "What We Know and Don't Know About Online Word-of-Mouth: A Review and Synthesis of the Literature," *Journal of Interactive Marketing*, 28 (3), 167–83.

Kramer, Mark A. (2007), "Self-Selection Bias in Reputation Systems," in *Trust Management*. *IFIPTM 2007. IFIP International Federation for Information Processing*, Vol. 238, S. Etalle and S. Marsh, eds. Boston: Springer.

Lebow, Jay L. (1982), "Consumer Satisfaction with Mental Health Treatment," *Psychological Bulletin*, 91 (2), 85–86.

Li, Neng, Suguo Du, Haizhong Zheng, Minhui Xue, and Haojin Zhu (2018), "Fake Reviews Tell No Tales? Dissecting Click Farming in Content-Generated Social Networks," *China Communications*, 15 (4), 98–109.

Li, Xinxin and Lorin M. Hitt (2008), "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research*, 19 (4), 456–74.

Liu, Yong (2006), "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, 70 (3), 74–89.

Luca, Michael and Georgios Zervas (2016), "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62 (12), 3412–27.

Marlin, Benjamin M., Richard S. Zemel, Sam Roweis, and Malcolm Slaney (2012), "Collaborative Filtering and the Missing at Random Assumption," *arXiv preprint arXiv*: *1206.5267*.

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

Minnema, Alec, Tammo H. Bijmolt, Sonja Gensler, and Thorsten Wiesel (2016), "To Keep or Not to Keep: Effects of Online Customer Reviews on Product Returns," *Journal of Retailing*, 92 (3), 253–67.

Mittal, Banwari and Walfried M. Lassar (1998), "Why Do Customers Switch? The Dynamics of Satisfaction Versus Loyalty," *Journal of Services Marketing*, 12 (3), 177–94.

Mittal, Vikas and Wagner A. Kamakura (2001), "Satisfaction, Repurchase Intent, and Repurchase Behavior: Investigating the Moderating Effect of Customer Characteristics," *Journal of Marketing Research*, 38 (1), 131–42.

Moe, Wendy W., Oded Netzer, and David A. Schweidel (2017), "Social Media and User Generated Content Analysis," in *Handbook of Marketing Decision Models*, Berend Wierenga and Ralf van der Lans, eds. Berlin: Springer.

Moe, Wendy W. and David A. Schweidel (2012), "Online Product Opinions: Incidence, Evaluation, and Evolution," *Marketing Science*, 31 (3), 372–86.

Moe, Wendy W. and Michael Trusov (2011), "The Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research*, 48 (3), 444–56.

Moors, Guy, Natalia D. Kieruj, and Jeroen K. Vermunt (2014), "The Effect of Labeling and Numbering of Response Scales on the Likelihood of Response Bias," *Sociological Methodology*, 44 (1), 369–99.

Muchnik, Lev, Sinan Aral, and Sean J. Taylor (2013), "Social Influence Bias: A Randomized Experiment," *Science*, 341 (6146), 647–51.

Mudambi, Susan M. and David Schuff (2010), "What Makes a Helpful Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly*, 34 (1), 185–200.

Naylor, Gillian and Susan B. Kleiser (2000), "Negative Versus Positive Word-of-Mouth: An Exception to the Rule," *Journal of Satisfaction, Dissatisfaction and Complaining Behavior*, 13 (1), 26–36.

Peterson, Robert A. and William R. Wilson (1992), "Measuring Customer Satisfaction: Fact and Artifact," *Journal of the Academy of Marketing Science*, 20 (1), 61–71.

Saleh, Khalid (2015), "The Importance of Online Customer Reviews," *Invesp* (accessed July 9, 2020), www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic.

Schlosser, Ann E. (2005), "Posting Versus Lurking: Communicating in a Multiple Audience Context," *Journal of Consumer Research*, 32 (2), 260–65.

Sharpe Wessling, Kathryn, Joel Huber, and Oded Netzer (2017), "MTurk Character Misrepresentation: Assessment and Solutions," *Journal of Consumer Research*, 44 (1), 211–30.

Silverman, George (1997), "How to Harness the Awesome Power of Word of Mouth," *Direct Marketing*, 60 (7), 32–37.

Skowronski, John J. and Donal E. Carlston (1987), "Social Judgment and Social Memory: The Role of Cue Diagnosticity in Negativity, Positivity, and Extremity Biases," *Journal of Personality and Social Psychology*, 52 (4), 689–98.

Sridhar, Shrihari and Raji Srinivasan (2012), "Social Influence Effects in Online Product Ratings," *Journal of Marketing*, 76 (5), 70–88.

Weijters, Bert, Elke Cabooter, and Niels Schillewaert (2010), "The Effect of Rating Scale format on Response Styles: The Number of Response Categories and Response Category Labels," *International Journal of Research in Marketing*, 27 (3), 236–47.

Wolff-Mann, Ethan (2016), "Here's Everything Wrong with Online Reviews and How to Fix Them," Time Money, Available at: http://time.com/money/page/online-reviews-trust-fix/.

You, Ya, Gautham G. Vadakkepatt, and Amit M. Joshi (2015), "A Meta-Analysis of Electronic Word-of-Mouth Elasticity," *Journal of Marketing*, 79 (2), 19–39.

Zervas, Georgios, Davide Proserpio, and John W. Byers (2015), "A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average," SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2554500.

Zhang, Yinlong, Lawrence Feick, and Vikas Mittal (2014), "How Males and Females Differ in Their Likelihood of Transmitting Negative Word of Mouth," *Journal of Consumer Research*, 40 (6), 1097–108.