REVIEWING EXPERTS' RESTRAINT FROM EXTREMES AND ITS IMPACT ON SERVICE PROVIDERS

Peter Nguyen, Xin (Shane) Wang, Xi Li, and June Cotte

Author affiliations

Peter Nguyen
Assistant Professor of Marketing
Miami University
Oxford, OH, USA, 45056
Email: pnguyen@miamioh.edu

Xin (Shane) Wang
Associate Professor of Marketing and Statistics
MBA '80 Faculty Fellow
Ivey Business School, Western University
London, Ontario, Canada, N6G 0N1
Email: xwang@ivey.ca

Xi Li
Assistant Professor of Marketing
City University of Hong Kong
Kowloon Tong, Hong Kong
Email: xili44@cityu.edu.hk

June Cotte
Professor of Marketing
Scott & Melissa Beattie Professorship in Marketing
Ivey Business School at Western University
London, Ontario, Canada, N6G 0N1
Email: jcotte@ivey.ca

This article is based on the first author's dissertation.

Accepted to the *Journal of Consumer Research* June 24, 2020.

REVIEWING EXPERTS' RESTRAINT FROM EXTREMES AND ITS IMPACT ON SERVICE PROVIDERS

Abstract

This research investigates reviewing experts on online review platforms. The main hypothesis is that greater expertise in generating reviews leads to greater restraint from extreme summary evaluations. The authors argue that greater experience generating reviews facilitates processing and elaboration, and enhances the number of attributes implicitly considered in evaluations, which reduces the likelihood of assigning extreme summary ratings. This restraint-of-expertise hypothesis is tested across different review platforms (TripAdvisor, Qunar, and Yelp), shown for both assigned ratings and review text sentiment, and demonstrated both between (experts vs. novices) and within reviewers (expert vs. pre-expert). Two experiments replicate the main effect and provide support for the attributes-based explanation. Field studies demonstrate two major consequences of the restraint-of-expertise effect. (i) Reviewing experts (vs. novices), as a whole, have less impact on the aggregate valence metric, which is known to affect page-rank and consumer consideration. (ii) Experts systematically benefit and harm service providers with their ratings. For service providers that generally provide mediocre (excellent) experiences, reviewing experts assign significantly higher (lower) ratings than novices. This research provides important caveats to the existing marketing practice of service providers incentivizing reviewing experts, and provides strategic implications for how platforms should adopt rating scales and aggregate ratings.

Keyword: Online word-of-mouth, Expertise, User rating average, Platform strategy, Text analysis, sentiment analysis

REVIEWING EXPERTS' RESTRAINT FROM EXTREMES AND ITS IMPACT ON SERVICE PROVIDERS

Consumers rely on the opinions and recommendations of others. Many of these recommendations have come from expert professionals (e.g., sommeliers, movie critics). Over the past couple of decades, the world has seen the rise of online reviews, where consumers not only rely on other consumers' experiences, but also share their own. Online review platforms now recognize their top users as reviewing 'experts'. For example, Yelp has its 'Elite' status, TripAdvisor has its 'Contributor Level', Google has its 'Local Guide' badges, and Amazon has its 'Amazon Vine Program.' Given that consumers are increasingly both sharing and consuming reviews, understanding the nature of reviewing experts has become an important topic in consumer research.

The study of online reviewing experts is particularly important for *service providers*, such as hotels and restaurants. Many businesses incentivize, by quite literally wining and dining, online reviewing experts, in order to get them to write high quality reviews for the business (Chakrabarti 2013; Stone 2014). The underlying assumption is that having reviews written by reviewing experts ultimately helps the business. Therefore, a very important managerial question is whether this assumption is (always) true.

Understanding online reviewing experts is also critical for *review platforms*, such as TripAdvisor and Yelp. A major goal of online review platforms is to (accurately) collect the experiences of past customers and present that information to prospective review-seeking customers. Given that many review platforms can and do distinguish amongst their users, understanding differences between reviewing experts and novices can shape how various aspects of the platform are designed in order to more accurately capture and display past customer experiences.

In our research, we use the term *reviewing experts* to refer to users on review platforms that the platforms designate as experts, such as Yelp's 'Elite' reviewers and TripAdvisor's top 'Contributor Level' reviewers. Although there are some differences among how platforms designate their reviewing experts, some common criteria include having generated a high *quantity* of reviews, and generating reviews that are of high *quality*, where quality can be assessed across a number of dimensions, such as degree of elaboration (review length) and review favorability judged by readers (number of 'Like' votes the review receives). Thus, we define a reviewing expert as a reviewer who has a high number of higher than average quality reviews, whose reviews are more likely to be judged as favorable by readers.

Although our construct of reviewing expertise is similar to the traditional construct of expertise, as defined in the literature (e.g., Alba and Hutchinson 1987; Johnson and Mervis 1997), there are some important differences. First, due to the practical nature of how review platforms designate their reviewing experts, we place greater emphasis in defining reviewing expertise on its task-related dimension (Proposition 1 from Alba and Hutchinson 1987) than its knowledge-related dimension (Proposition 2 from Alba and Hutchinson 1987). Hence, to be clear, we are not per se studying individuals with a high degree of knowledge about the hotel/restaurant industry, although this may be true of many of the individuals that we are examining. Instead, our focus is on individuals who write lots of reviews that are of above average quality, and as a result, these consumers have been deemed as "experts" by the review platform.

Second, the operationalization of online reviewing experts does not involve a standardized qualifying test, such as in the case with medical doctors and sommeliers. Instead, due to their large userbase and the high degree of variance in how users engage on the platforms, online

review platforms adopt quick and scalable approaches for identifying reviewing experts. For example, TripAdvisor uses a transparent point-based system (Tripadvisor 2020), and Yelp adopts a user nomination system (Yelp Support Center 2020). We acknowledge the imperfection in a quick and scalable approach in designating reviewing expertise; however, as we show in our research using data from three different review platforms, the scalable approaches adopted by review platforms are reasonable proxies for capturing expertise.

As the focus of our research is on the relationship between reviewing expertise and review evaluations, while we do include some measures of consumer perceptions (e.g., 'Like', 'Helpful' and 'Useful' votes), it is not our intention to fully elucidate the perceptions of review-*reading* consumers of expert-generated reviews, but to focus on the effects of reviewing expertise on rating evaluations.

Our research makes four key contributions. First, we bridge the gap between the topic of online reviewing expertise and the more general literature on expertise (e.g., Alba and Hutchinson 1987). We provide empirical evidence that online reviewing "experts," as designated by many online review platforms, largely exhibit features of expertise, including a greater degree of elaboration, and greater category knowledge.

Second, we explain the relationship between reviewing expertise and evaluative rating patterns. Extant research shows that compared to novices, experts are generally more critical (i.e., more negative) in their evaluations (e.g., Amabile 1983; Mollick and Nanda 2016; Schlosser 2005; Zhang, Zhang, and Yang 2016). Across our three field studies, we find this to be true, but only when reviewing experts evaluate service providers that generally provide excellent experiences (i.e., above 4.0 stars). However, when experts evaluate service providers that generally provide mediocre experiences (i.e., 3.5 stars or below), the opposite is true – novices

are more critical than experts. Our overarching explanation is that experts (vs. novices) are more restrained from extreme summary ratings (i.e., our *restraint-of-expertise* hypothesis). Our explanation is consistent with past research on in- (vs. out-) groups (Linville and Jones 1980; Linville 1982), which explains that people have more complex cognitive representation of members of their own groups than those of other groups, and as a consequence, tend to evaluate members of their in- (vs. out-) group as less extreme.

Third, our research contributes to discussions on the observed extreme (J-shaped) rating distribution in online reviews (i.e., most reviewers assign 5-star ratings, some 1-star ratings, and few ratings fall in between). Much of the attribution for this observed pattern is the reviewer's motivation to generate reviews (a self-selection bias; Hu, Pavlou, and Zhang 2009; Schoenmüller, Netzer, and Stahl 2019). For example, consumers are more likely to write reviews when their experiences are really good or really bad. We agree that self-selection plays an important role in influencing the extent to which extreme rating distributions are observed. However, our research points to another important factor, *reviewing expertise*. Novice evaluators generally think in a binary/polarizing fashion (Linville 1982; Rozin, Ashmore and Markwith 1996), but we find that as they gain greater experience generating reviews, they (implicitly) consider more attributes in their evaluations, and in turn, provide summary ratings that are more restrained from the extremes. Thus, reviewing expertise for a particular product/service influences the extent to which extreme rating distributions are observed.

Fourth, although much of the extant research on online reviews provides support for the consequences of the aggregate valence metric, such as consumer choice and firm sales (Floyd et al. 2014; Luca 2016), little to nothing is known about its *antecedents* (Dai et al. 2018). Our research uncovers one such antecedent. We show that based on their rating approach, reviewing

experts (vs. novices), as a whole, play a lesser role in shifting the aggregate valence metric. Our findings complement and refine the conventional notion that expert recommendations highly affect consumer choice (Biswas, Biswas, and Das 2006; Chocarro and Cortiñas 2013; Karmarker and Tormala 2009). Although the actual review content generated by experts is generally favored by consumers (Racherla and Friske 2012; Zhang, Zhang, and Yang 2016), the attenuated impact experts have on the aggregate valence metric means that reviewing experts (vs. novices) have a less important role in shaping the service providers that consumers will consider before reading individual reviews (Ghose, Ipeirotis, and Li 2012; Vermeulen and Daphne 2009).

Our results have two important managerial implications. First, we provide caveats to the common business practice of active solicitation of reviewing experts (Stone 2014). We delineate when and how reviewing experts benefit and harm service providers, in terms of systematically raising and lowering the aggregate valence metric. Second, our research brings to light the issue of adopting ratings scales with the same granularity for experts and novices, and the problem with combining expert and novice ratings to form aggregate valence metrics. We recommend review platforms adopt different rating scales for their expert and novice users (using a more granular scale for their experts), and present different aggregate valence metrics for ratings by these two groups. An in-depth discussion on the managerial implications and our recommendations is provided in the discussion section.

The rest of the paper is organized as follows. We first present a review of the background literature on online reviews and reviewing expertise, followed by our proposed hypotheses. Next, we present our five studies (three field studies and two randomized controlled experiments). Lastly, we discuss our main findings and provide managerial implications for service providers and rating platforms.

Overview of the Literature

Online peer reviews have been an important topic in marketing over the last decade. Given the information asymmetry between firms and consumers (Mishra, Hedide, and Cort 1998), online reviews play a major role in reducing the information gap and shaping consumer choice (Hu, Liu, and Zhang 2008). For instance, marketing researchers have demonstrated the impact of online peer reviews on consumer choice (Luca 2016) and firm sales (Floyd et al. 2014).

Much of the existing research on online reviews can be categorized, based on the level of analysis, into two groups: aggregate (e.g., Chevalier and Mayzlin 2006; Moe and Trusov 2011; Sonnier, McAlister, and Rutz 2011) and individual-level reviews (e.g., Liu and Park 2015; Packard and Berger 2017; Yin, Bond, and Zhang 2017). In aggregate-level review research, the unit of analysis is at the level of the product/service, where individual reviews are grouped across each product/service to form aggregate metrics. A major finding in this area is that aggregate metrics, such as valence and volume, are predictive of firm sales (Babić Rosario et al. 2016; Floyd et al. 2014; You, Vadakkepath, and Joshi 2015). Aggregate metrics are important to service providers because they influence the page on which service providers appear on review platforms (Ghose, Ipeirotis, and Li 2012), and are used by consumers to form their consideration set before reading individual reviews (Dai et al. 2018; Fisher, Newman, and Dhar 2018). Although research has been conducted on the predictive nature of aggregate metrics, little is known about their antecedents. For instance, are there specific types of reviewers that tend to shift the existing aggregate valence metrics more, assigning ratings that are more distant from the current user rating averages? If so, who? In which direction? Studying the antecedents of the valence metric is important because it provides practitioners and researchers with clues regarding factors that affect the products/services consumers consider.

In individual-level online review research, the unit of analysis is the individual review. Researchers examine how consumer opinions are influenced by review characteristics, such as star rating, review length, and mobile-generated review labels (Grewal and Stephen 2019; Liu and Park 2015; Mudambi and Schuff 2010; Peng et al. 2014), measures of review content, such as readability, expressed emotions, and implicit/explicit endorsements (Korfiatis, García-Bariocanal, and Sánchez-Alonso 2012; Packard and Berger 2017; Yin, Bond, and Zhang 2017), and reviewer characteristics, such as reputation and disclosure of identity (Liu and Park 2015; Racherla and Friske 2012). Given that many review platforms can and do distinguish amongst their users, it is surprising that we actually know little about reviewing expertise.

Reviewing Experts

A few studies have been published on reviewing experts (e.g., Liu and Park 2015; Zhang, Zhang and Yang 2016). Researchers have operationalized expertise in terms of number of past reviews generated; no overarching conceptual definition has yet been provided, and no empirical link has yet been tested, between online reviewing "expertise" and the traditional literature on expertise (e.g., Alba Hutchinson 1987). There are some indications that reviewing experts have more source credibility than their novice counterparts (Racherla and Friske 2012; Vermeulen and Seegers 2009; Zhang, Zhang, and Yang 2016), however, the relationship between online reviewing "experts" and other features of expertise, such as degree of elaboration and degree of domain-specific knowledge (Alba and Hutchinson 1987), remain to be empirically tested.

Researchers have demonstrated a lack of consistency between expert judgments and lay people's opinions (de Langhe, Fernbach and Lichtenstein 2016; Holbrook 1999), even at the aggregate level (Dai et al. 2018). Relatedly, many research studies show that in their evaluations, experts are generally more critical than novices (Amabile 1983; Mollick and Nanda 2016;

Zhang, Zhang, and Yang 2016). Given the differences between expert and novice ratings, uncovering boundary conditions for the (in)consistency of expert and novice ratings can provide valuable insight into how experts and novices make evaluations and show how businesses are benefited/harmed by expert (vs. novice) evaluations.

Extreme Rating Distribution

One of the key observations from research on online reviews is the extreme (J-shaped) rating distribution (Hu, Pavlou, and Zhang 2006, 2009); that is, most reviewers assign 5-star ratings, some assign 1-star ratings, and there are few ratings in between. The key explanation for this finding is related to reviewers' self-selection for generating reviews (Schoenmüller, Netzer, and Stahl 2019). That is, consumers are more likely to write and post reviews when experiences are extreme. The empirical evidence in support of this claim is from the observed negative correlation between the number of past reviews generated by a reviewer and the assignment of extreme ratings; reviewers who rarely post reviews are more likely to assign 1- and 5-star ratings. However, it is unclear whether other factors, aside from self-selection in reviewing, might contribute to explaining this negative correlation. In our research, we investigate the role of *reviewing expertise* in the observation of the extreme rating distribution.

There is evidence that novices, by their nature, are more polarizing/dichotomous in their evaluations (e.g., Rozin, Ashmore and Markwith 1996). Research on in-group versus out-group evaluations and political cognition have shown that people evaluate out-group members more extremely than in-group members (Fiske, Kinder, and Larter 1983; Linville and Jones 1980). The explanation is that people have more complex cognitive representation of members of their own group than those of other groups, and that the less complex a person's representation of stimuli from a given domain, the more extreme will be the person's evaluations of stimuli from that

domain (Linville 1982). Experimental evidence supports this explanation (Linville 1982; Linville and Jones 1980), suggesting a causal link between complex representation of stimuli in a domain and less extreme ratings. Given that experts, by their nature, have a more refined cognitive structure for a particular category (Alba and Hutchinson 1987), we might expect that the degree of observed extreme ratings, in part, be explained by (lack of) reviewing expertise. We elaborate further in the next section.

Theory and Hypotheses

Repetition and Expertise

A major question regarding online reviewing experts, such as Yelp's 'Elite' reviewers, is whether they actually display features of expertise (e.g., Alba and Hutchinson 1987; Harmon and Coney 1982). To address this question, a clear understanding of how *review platforms* operationally define their reviewing experts is required. To define their reviewing experts, review platforms generally assess their reviewers across a number of dimensions, including the number of past reviews generated and inclusions of photo/video. For most review platforms, such as Qunar and TripAdvisor, the designation of expertise level is done automatically using a transparent point-based system, where reviewers receive points for their contribution to the platform (e.g., generating a review, including photos/videos in their review). Reaching milestones moves reviewers up in designated expertise level (Tripadvisor 2020). For other platforms, such as Yelp, various aspects of contribution to the platform are also taken into consideration, but the designation of expertise is done by humans (e.g., other reviewers on the platform nominate a reviewer for the expert designation and a 'Community Manager' decides on whether or not that reviewer receives the official expertise badge; Yelp Support Center 2020).

Across most review platforms, a common criterion of 'expertise' is generating lots of reviews. Extant research on expertise highlights the importance of practice/repetition in the development of expertise (Alba and Hutchinson 1987; Hintzman 1976). According to Alba and Hutchinson (1987), repetition improves task performance by reducing cognitive effort, refines domain-related cognitive-structure, and enhances the ability to elaborate. Therefore, given that most review platforms adopt some measure of quantity of reviews in their expertise designation, we expect, and empirically validate, that platform-defined reviewing experts actually do display quality-based expertise features, such as greater review elaboration, greater domain-specific knowledge, and greater review favorability as rated by readers. We first need to establish that our conceptual definition of reviewing expertise aligns with the operational definition of expertise used by the various review platforms. So before testing our theory, in our studies we establish this equivalence: Platform-defined experts (a) elaborate more, (b) display greater domain-knowledge, and (c) generate reviews deemed more favorable to readers.

Expertise and Rating Patterns

An important research question about reviewing experts is how expertise in generating reviews affects rating evaluations, if at all. Given that repetition of reviewing is a common criterion in operationalizing reviewing expertise, and that repetition facilitates processing (Einhorn and Hogarth 1981; Hoyer 1984) and elaboration (Mandler and Johnson 1981), we predict that with greater experience in generating reviews, reviewers come to implicitly consider more domain-specific attributes (e.g., *price*, *environment*, *location*, and *service*) in their evaluations (Johnson and Mervis 1997). Because product/service summary ratings are generally derived from (implicit) ratings across considered attributes (Hong and Wyer 1989; Nowlis and Simonson 1996), and due to the regression towards the mean principle (Stigler 1997), we predict

that the consideration of larger numbers of attributes in evaluations reduces the likelihood of assigning extreme summary ratings. In other words, we acknowledge that the assignment of extreme ratings can and do occur across all reviewers; however, we argue that the assignment of extreme ratings generally requires that the service provider perform consistently excellent, or consistently terrible, across all attributes considered by the reviewer, which is a lot less likely when reviewers consider more attributes in their evaluations.

H1 (The restraint-of-expertise hypothesis): Greater expertise in generating reviews leads to greater restraint from extremes in summary evaluations.

H2: The restraint-of-expertise effect (H1) is driven by the number of attributes considered in the evaluation.

Downstream Consequences of the Restraint-of-Expertise Hypothesis

Although Hypotheses 1 and 2 may be of particular interest to consumer researchers, practitioners are more concerned with the 'so-what' question. We predict two important downstream consequences that might arise as a result of the restraint-of-expertise hypothesis. The downstream consequences deal with (i) the shifting of the aggregate valence metric and (ii) the relative ratings between experts and novices.

As we have mentioned, aggregate valence metrics are predictive of firm sales (Babić Rosario et al. 2016; You, Vadakkepath, and Joshi 2015), influence where service providers appear on review platforms (Ghose, Ipeirotis, and Li 2012) and are used by consumers to form consideration sets (Luca 2016; Vermeulen and Daphne 2009). This metric is clearly important to marketers. Because rating averages, by their nature, are generally skewed from extreme values, we expect that as a consequence of their less polarizing rating approach:

H3: Reviewing experts (vs. novices) play a lesser role in shifting the aggregate valence metrics.

An important follow-up question to H3 is whether novices (vs. experts) shift the aggregate valence metric *randomly* (i.e., equally shifting it up and down, where the net movement of the aggregate valence metric is neutral) or *directionally* (i.e., shifting it up or down, where the net movement is positive or negative). We suspect novices' impact on the aggregate valence metrics is directional, and dependent on the general level of service by the service provider. Our idea is that based on the restraint-of-expertise hypothesis, relative to reviewing experts, novices adopt a more polarizing approach (i.e., an "I love it" vs. "I hate it" mentality). When presented with a positive experience, novice reviewers are a lot more likely to rate the experience as excellent (e.g., a rating of 5 on a 5-point scale) compared to expert reviewers, who are hesitant to give an extreme positive rating, given all the attributes they consider. Conversely, when presented with a negative experience, novice reviewers are more likely to rate the experience as terrible (a rating of 1) compared to expert reviewers, who are hesitant to give an extreme negative rating, as they consider multiple attributes of the experience. Therefore, we hypothesize:

H4: For service providers that generally provide mediocre (excellent) experiences, expert reviewers assign higher (lower) ratings than novice reviewers.

Overview of Studies

In this section, we present five research studies (three field studies and two experiments) investigating our hypotheses. An overview of the three field studies can found in **Table 1**. It is important to note that we collected and analyzed two types of review data: (i) reviews based on *service providers* and (ii) reviews based on *reviewers*. The by-service-provider (by-reviewer) data consists of all the reviews on a number of service providers (by a number of reviewers).

Both types of data are necessary to address alternative explanations to the restraint-of-expertise effect that are related to reviewers' selection of service providers and reviewers' self-selection for writing reviews.

* * * TABLE 1 ABOUT HERE * * *

First, it is conceivable that the fact that reviewing experts (vs. novices) are more restrained in their ratings might simply be because experts are more likely to visit and write about a wider range of service providers, including those that provide mediocre service levels. We address and mitigate this concern of reviewers' selection of service providers with our by-service-provider data, where the specific service provider selected by the reviewer is kept constant.

Second, researchers have shown that the extreme rating distribution observed in online reviews is largely due to reviewer's motivation to write reviews (a self-selection bias; Hu et al. 2009; Schoenmüller et al. 2019). Hence, it is plausible that experts (vs. novices) are more restrained just because novices, who do not write many reviews, only write reviews when experiences are really good or really bad. To mitigate the concern that the restraint-of-expertise effect might just be explained by a self-selection bias, we collected by-reviewer data, where we analyze how the review ratings and content of reviewers change as they gain greater experience generating reviews.

In additional to the field data, which provide generalizability – to the real world and across platforms – of the restraint-of-expertise hypothesis and its downstream consequences, two experiments were conducted to strengthen our claim regarding the causal inference and attributes-based explanation for our phenomenon of interest.

Consistent with our definition of reviewing expertise, across all our studies, we measure/manipulate reviewing expertise in terms of quantity and/or quality. In the three field

studies – with data from Qunar, TripAdvisor, and Yelp – we operationalize reviewing expertise in terms of the platform's designation of expertise, which, as explained, is measured across a number of dimensions including the number of past reviews generated (quantity), whether photos/videos were included in reviews, and the number of 'Like' votes the reviews receive (quality). In the experimental studies, we manipulate reviewing expertise in terms of rating repetition (~quantity) and number of attributes considered (~quality).

As discussed in the introduction, due to the practical nature of how review platforms designate their reviewing experts, we place greater emphasis in defining reviewing expertise on its task-related dimension (Proposition 1 from Alba and Hutchinson 1987) than its knowledge-related dimension (Proposition 2). Relatedly, in our experimental studies, the objective was to manipulate the task (vs. knowledge) dimension of expertise, in a manner that is similar to how online review platforms operationalize their reviewing experts.

Study 1: Qunar (Field Data)

Purpose. The main goal of Study 1 is to examine the relationship between reviewing expertise and assigned rating patterns.

Variables and Analyses. In Study 1, we scrape and analyze over 125,000 online reviews of hotels on Qunar.com, a major online travel review platform in China (see **Table 1** for description of dataset; see **Table 2** for variable list; see **Table 3** for summary statistics of variables).

* * * TABLES 2 & 3 ABOUT HERE * * *

The main independent variable of interest is *reviewing expertise*, which is conceptually defined in terms of quantity and quality. In this study, we operationalize reviewing expertise the way the platform does; that is, based on Qunar's platform-defined *1-7 Expertise Level*. As previously mentioned, Qunar measures its reviewing experts using a transparent point system,

where points are predominantly assigned for the quantity of reviews generated by a reviewer. We used the natural logarithm of Qunar's *1-7 Expertise Level*, i.e., ln(*Expertise_level*), in our analysis to normalize its distribution. Throughout the analyses, we provide descriptive statistics for the first two Expertise Levels, levels 1 and 2, and the last two Expertise Levels, levels 6 and 7.

The main dependent variables of interest are rating polarity and impact of rating on the aggregate valence metrics. *Rating polarity* is operationalized as the distance of the reviewer's assigned rating from the midpoint of the scale. In the case of Qunar, which adopts a 5-point rating scale, rating polarity is measured as the absolute value of the assigned rating subtracted by the scale-midpoint value of 3; i.e., $|Rating - 3|$.

Impact of rating on the aggregate valence metric is the degree to which an assigned rating shifts the user rating average. It is measured as the absolute difference between a reviewer's assigned star rating and the service provider's average consensus rating at the point in time the reviewer is assigning the rating; this is a dynamic variable. For example, if a hotel's average rating is 4.2 and then a reviewer gives the hotel a rating of 3 out of 5, then the rating-average distance for this review is 1.2. For robustness of measurement, we operationalize impact of ratings on both the *moving* valence metric (based on most recent 20 reviews at time of assigning the rating) and the *cumulative* valence metric (based on all past reviews at time of assigning the rating).

Because there are multiple reviews of each hotel, that is, the reviews are nested within hotels, we conduct our main analyses with linear mixed-effects regressions, with maximum likelihood estimation. Included in the analyses are a number of control variables, including hotel ID (as a random effect, *ServiceProvider*), date of review post (converted to number of months

from date of review scraping, *MonthsAgo*), expertise level of the prior reviewer posting about the service provider (to control for some interdependencies amongst reviewers, *PriorReviewer*), and purpose of travel (transformed to five dummy variables, *Purpose*).

Level 1: $RatingPolarity_{ij} = \beta_{0j} + \beta_1 \ln(ExpertiseLevel)_{ij} + \beta_2 MonthsAgo_{ij} + \beta_3 \ln(PriorReviewer)_j + \beta_{4\text{-}8} Purpose_{ij} + \varepsilon_{ij}$

Level 2: $\beta_{0j} = \gamma_0 + \gamma_1 ServiceProvider_j + \mu_j$

Results: (i) Platform-Defined Reviewing Expert. To establish whether Qunar's platform-defined reviewing 'expert' designation is consistent with the literature-defined concept of expertise (Alba and Hutchinson 1987; Harmon and Coney 1982), we examine how various quality-based expertise features of reviews vary as a function of Qunar's platform-defined expertise levels. We find that reviewers higher on Qunar's *1-7 Expertise Level* (i) have a higher degree of elaboration in their reviews ($M_{Levels\_1\_2}$ = 74 vs. $M_{Levels\_6\_7}$ = 1611 Chinese characters per review, r = .13, p < .001; robustness test of only reviews within 3 standard deviations of the review length mean: $M_{Levels\_1\_2}$ = 66 vs. $M_{Levels\_6\_7}$ = 243 Chinese characters per review, r = .08, p < .001), and (ii) generate reviews that are deemed more favorable by readers ($M_{Levels\_1\_2}$ = 0.3 vs. $M_{Levels\_6\_7}$ = 6.2 average 'Like' votes per review post, r = .07, p < .001; robustness test of only reviews with at least 1 'Like' vote: $M_{Levels\_1\_2}$ = 2.9 vs. $M_{Levels\_6\_7}$ = 8.5 average 'Like' votes per review post, r = .18, p < .001). Our conceptual definition of reviewing expertise aligns with the platform's operational definition of expertise.

(ii) Expertise and Rating Evaluations (H1). Next, we test the relationship between reviewing expertise and rating polarity. Consistent with H1, results from our linear mixed-effects regression model show that reviewers higher on Qunar's *1-7 Expertise Levels* demonstrate greater restraint from extremes in their ratings ($M_{Levels\_1\_2}$ = 1.62 vs. $M_{Levels\_6\_7}$ = 1.37 average distance away from midpoint of the five-point rating scale; β = -0.09, t(125917) = -23.43, p < .001, *Cohen's* f^2 = *0.066*; see **Figure 1A**). To test whether this observed relationship between reviewing expertise

and decreased rating polarity might just be due to reviewers who rarely post, we reran our analysis using only reviews by reviewers who have posted at least 10 and 20 reviews prior. Our results hold (at least 10 reviews: $\beta = -0.14$, $t(19860) = -8.09$, $p < .001$, $f^2 = 0.057$; at least 20 reviews: $\beta = -0.14$, $t(6200) = -3.41$, $p < .001$, $f^2 = 0.043$).

* * * FIGURE 1 ABOUT HERE * * *

We relax our parametric assumption about the rating polarity dependent variable by conducting a more conservative test, ordered logistic regression (using *polr()* function in the *MASS* package in R; Ripley et al. 2013); our restraint-of-expertise results hold ($\beta = -0.33$, $t = -24.55$, $p < .001$). As a robustness check of the measurement of the dependent variable, rating polarity, we conduct an analysis comparing the dispersion of the star ratings assigned by experts and novices. Results from Bartlett's test of homogeneity of variances show that the variance of ratings by experts ($SD_{Level\_6\_7} = 0.68$) is significantly lower than the variance of ratings by novices ($SD_{Level\_1\_2} = 0.91$; $K^2 = 57.50$, $p < .001$).

Our explanation for the restraint-of-expertise effect is based on attributes implicitly considered by reviewers when making their overall rating evaluation (H2). Later, in our English-language review data, we algorithmically detect and count the number of category-related nouns mentioned in the review itself as a measure of the number of considered attributes. In the Qunar review data, due to limitations in analyzing Chinese text, we are unable to extract the specific attributes mentioned in the reviews. We do, however, use review length, in Chinese characters, as a proxy for the number of considered attributes. Using mediation analysis in R (*mediation* R package, Tingley et al. 2014), we test the mediating role of review length on the relationship between reviewing expertise and restrained ratings. Conducting 1000 iterations, the number-of-considered-attributes proxy, review length, was found to be a significant mediator (-0.0178, 95%

CI: -0.0192 to -0.0164), accounting for 19.4% of the covariance between reviewing expertise and decreased rating polarity. This finding provides support for our claim that experts consider more attributes, which leads to a less extreme, or restrained, overall evaluation.

(iii) Impact of expertise on shifting the aggregate valence metric (H3). Next, we test the impact of expertise on the aggregate valence metric. Consistent with H3, the results from our mixed-effects model demonstrate a significant negative effect of reviewing expertise on the aggregate valence metric – both in terms of the moving valence metric (i.e., difference of assigned rating from most recent 20 reviews on service provider at time of reviewer posting; $M_{Level\_1\_2} = 0.63$ vs. $M_{Level\_6\_7} = 0.56$; $\beta = -0.48$, $t(124870) = -8.90$, $p < .001$, $f^2 = 0.025$) and the cumulative valence metric (i.e., difference of assigned rating from all past reviews on service provider at time of reviewer posting; $M_{Level\_1\_2} = 0.67$ vs. $M_{Level\_6\_7} = 0.58$; $\beta = -0.50$, $t(125916) = -5.29$, $p < .001$, $f^2 = 0.015$). In other words, we find that compared to their novice counterparts, reviewing experts shift service providers' aggregate valence metric less.

Conclusions. In Study 1, collecting and analyzing Qunar hotel review data, we demonstrate that although Qunar adopts a predominantly quantity-based expertise designation, their platform-defined reviewing experts by and large display quality-based expertise as well, in terms of greater review elaboration and greater reader-assessed review favorability. We show that reviewing experts (vs. novices) adopt a less polarizing rating approach (H1), which appears to be in part driven by how many words/characters (our proxy for attributes) they use in their evaluations (H2). As a consequence, reviewing experts shift aggregate valence metrics less (H3), which is managerially important, as valence metrics affect page-rank (Ghose, Ipeirotis, and Li 2012) and consumer consideration (Luca 2016; Vermeulen and Daphne 2009).

An advantage of collecting and analyzing the field data is the ability to draw claims about the generalizability of observed findings in the real world. However, a major drawback concerns the nature of the relationship between the variables of interest, in our case, reviewing expertise and less polarizing rating evaluations. As previously mentioned, extant literature on online reviews suggest that the observed extreme rating distribution on many online review platforms is largely attributed to the reviewers' motivation to write reviews (i.e., a self-selection bias; Hu, Pavlou, and Zhang 2009; Schoenmuller, Netzer, and Stahl 2018); reviewers are more likely to write reviews when experiences are really good or really bad. Hence, this begs the question, is the observed phenomenon driven purely by a self-selection bias? Or is the relationship also causal in nature, such that as reviewers gain expertise, their reviews, both in terms of assigned ratings and review text sentiment, become more restrained?

We speculate that, to some degree, both a self-selection bias and a causal relationship are present in the relationship between reviewing expertise and decreased rating polarity. In the subsequent studies, we test and demonstrate the causal effect of reviewing expertise on less polarizing rating evaluations. We conduct experiments in Studies 2A and 2B, where we manipulate aspects of reviewing expertise – rating repetition and number of considered attributes – to test the effect of reviewing expertise on less polarizing rating evaluations. As mentioned, a major goal of the experimental studies is to manipulate, and observe the consequences of, the task (vs. knowledge) dimension of reviewing expertise, in a manner that is similar to how online review platforms operationalize their reviewing experts. In Studies 3 and 4, analyzing TripAdvisor and Yelp reviews, we further test and provide evidence for the effect of reviewing expertise on less polarizing rating evaluations by tracking, intra-reviewer, how the polarity of assigned ratings and review text sentiment change as reviewers generate more reviews.

Study 2A: Priming an Aspect of Reviewing Expertise: Rating Repetition (Experiment)

Purpose. The purpose of Study 2A is to test the effect of reviewing expertise on the polarity of rating evaluations. A key criterion, across more-or-less all review platforms, in operationalizing reviewing experts is the number of past reviews generated. And research on expertise highlights the importance of practice/repetition in the development of expertise (Alba and Hutchinson 1987; Hintzman 1976). So, in Study 2A, we test whether a key aspect of reviewing expertise, as measured by review platforms, rating repetition (i.e., having assigned many reviews), affects the polarity of rating evaluations. Consistent with H1, we predict that greater rating repetition leads to ratings that are more restrained from extremes.

Design. The design of the experiment is a 2 rating repetition (high vs. low) x 2 description valence (positive vs. negative) between-subjects design. The outcome measure in the experiment is the assigned star rating, along a 5-point scale from 1-*Terrible* to 5-*Excellent* (see **Web Appendix A** for experimental stimuli).

Procedure. Online participants (N = 230, $\%_{female}$ = 55.3%, M_{Age} = 38.7, SD_{Age} = 13.5) on Amazon Mechanical Turk took part in the study. Participants were first asked to think about restaurants they have visited over the past year. Participants randomly assigned to the high (low) rating repetition condition were asked to write down five (two) of these restaurants, and then to rate each of these restaurant experiences along a 5-point rating scale, from 1-*Terrible* to 5-*Excellent*. Participants were then presented with a description of a positive or negative experience at a restaurant and then asked to assign a star rating for the description. Finally, as a control, participants were asked to report how often they write online reviews.

Results. As an attention check, we included a simple instruction item that required participants to select a particular multiple-choice option. Fifteen participants were removed because they failed the attention check, bringing the total number of participants to 215.

A two-way ANOVA revealed a significant main effect of description valence ($M_{positive}$ = 4.28 vs. $M_{negative}$ = 1.78; $F(1,211) = 764.14$, $p < .001$, $\eta_p^2 = .784$) and no main effect of rating repetition on assigned star rating (*ns*). As expected, the interaction between description valence and rating repetition on the assigned star rating was significant ($F(1,211) = 8.42$; $p = .004$, $\eta_p^2 = .038$; see **Figure 2A**). For robustness of analysis, we also conducted a two-way ANCOVA, including in the model the control variables age, gender, and frequency with which the participants write online restaurant reviews. Results are robust (main effect of description valence: $F(1,208) = 772.17$, $p < .001$, $\eta_p^2 = .788$; no main effect of rating repetition, *ns*; interaction between description valence and rating repetition: $F(1,208) = 6.29$, p = .013, $\eta_p^2 = .026$).

* * * FIGURE 2 ABOUT HERE * * *

A follow-up analysis shows that for the negative experience description, participants in the high rating repetition condition assigned significantly *higher* ratings ($M = 1.91$) than those in the low rating repetition condition ($M = 1.66$; $t(106) = 2.22$, $p = 0.028$, *Cohen's d* = 0.42). In contrast, for the positive experience description, participants in the high rating repetition condition assigned marginally *lower* ratings ($M = 4.13$) than those in the low rating repetition condition ($M = 4.40$; $t(84) = 2.22$, $p = 0.068$, $d = 0.38$). Next, we looked at the polarity rating variable, our main dependent variable. Consistent with our prediction, we find that participants in the high rating repetition condition (one dimension of reviewing expertise), assigned ratings that

were less polarizing (M = 1.21 average units from the midpoint of a five-point scale) than those in the low rating repetition condition (M = 1.37; $t(212) = 2.14$, $p = .033$, $d = 0.29$).

Conclusion. Using an experiment, we showed that priming a key aspect of reviewing expertise, rating repetition, reduces the polarity of ratings. This replicates the less polarizing rating approach favored by reviewing experts in the earlier Qunar field data. The parallel findings between our field data in Study 1 and our experiment data in Study 2A strengthen the conclusion of a causal relationship between reviewing expertise and restrained rating evaluations. To further test this causal relationship, in Study 2B, we conduct a similar experiment where we manipulate a different aspect related to reviewing expertise: number of considered attributes.

Study 2B: Priming an Aspect of Reviewing Expertise: Attribute Number (Experiment)

Purpose. The purpose of Study 2B is to further test the effect of reviewing expertise on the polarity of rating evaluations. Given our theorizing that reviewing experts consider more attributes in their evaluations, which contributes to the restraint-of-expertise effect, we test whether having participants consider a few or many attributes prior to assigning the summary rating affects the summary rating.

Interestingly, some platforms, like TripAdvisor, already have reviewers not only rate their overall experience, but also rate the experience along specific attributes. However, the attribute-level ratings are only done *after* the overall rating has been assigned. In Study 2B, ratings along attributes are done *before* assigning an overall rating. We test how the number of attributes considered might affect the overall rating. In alignment with H2, we hypothesize that considering a greater number of attributes when evaluating an experience, as experts are known to do, will lead to a more restrained summary rating.

Design. The design of the experiment is a 2 attribute number (high vs. low) x 2 experience valence (positive vs. negative). The outcome measure in the experiment is the assigned star rating, along a 5-star scale from 1-*Terrible* to 5-*Excellent* (see **Web Appendix B** for experimental stimuli).

Procedure. Online participants ($N = 240$, $\%_{female} = 60.2\%$, $M_{Age} = 37.4$, $SD_{Age} = 12.4$) on Amazon Mechanical Turk took part in the study. Participants were first randomly assigned to one of the experience valence conditions. Participants were asked to recall either a recent positive (or a recent negative) experience at a sit-down restaurant; they were asked to write the name of the restaurant, how long ago they visited the restaurant, and the number of times they have visited the restaurant.

Next, participants were randomly assigned to one of the two attribute number conditions. Participants were first asked to rate the recent restaurant experience across either six (high) or two (low) attributes, depending on the condition to which they were assigned (the selection of presented attributes was randomized). Then they were asked to give their summary star rating of the experience. All star ratings were assigned along a 5-star rating scale, from 1-*Terrible* to 5-*Excellent.* Finally, as a control, participants were asked to report how often they write online reviews in a month.

Results. As an attention check, we removed participants that were asked to report a positive (negative) restaurant experience, but reported an experience rating of 1-star (5-stars). This led to the removal of 24 of the 240 data points, bringing the total participant count to 216.

A two-way ANOVA revealed a significant main effect of experience valence ($M_{positive} = 4.23$ vs. $M_{negative} = 3.13$, $F(1,212) = 111.33$, $p < .001$, $\eta_p^2 = .344$), and no main effect of number of attributes on assigned star rating (*ns*). As predicted, the interaction between experience valence

and attribute number on the assigned star rating was significant ($F(1,212) = 5.32$, $p = .022$, $\eta_p^2 = .024$;see **Figure 2B**). For robustness of analyses, we also conducted a two-way ANCOVA, including in the model control variables age, gender, number of weeks ago participants visited the restaurant, number of times participant has visited the restaurant, and average number of times per month the participants writes online reviews. Results are robust (main effect of experience: $F(1,207) = 113.58$, $p < .001$, $\eta_p^2 = .354$; no main effect of number of attributes, ns; interaction between experience valence and attribute number: $F(1,207) = 4.49$, $p = .035$ $\eta_p^2 = .021$).

Following up on the interaction, we find that for the positive experience condition, participants primed to consider more attributes gave significantly lower individual summary ratings ($M_{6\_attributes} = 4.12$ vs. $M_{2\_attributes} = 4.36$; $t(111) = 2.19$, $p = .03$, *Cohen's d* = 0.40). For the negative experience condition, there was no significant effect of the number of attributes considered on assigned ratings ($M_{6\_attributes} = 3.24$ vs. $M_{2\_attributes} = 3.00$; *ns*).

To test the polarity of the individual summary ratings, we compare the variance of ratings by participants in the 6 (versus 2) attribute conditions. Results from Bartlett's test of homogeneity of variances show that the variance of summary ratings by participants in the 6-attribute condition ($SD_{6\_attributes} = 0.84$) is significantly lower than the variance of summary ratings by participants in the 2-attribute condition ($SD_{2\_attributes} = 1.06$; $K^2 = 5.86$; $p = .016$; see **Figure 2B**).

As a robustness of measurement, we also test the polarity of ratings based on the distance of the ratings from the average rating across all participants. We find that participants primed to consider more attributes gave significantly less polarizing ratings ($M_{6\_attributes} = 0.58$ vs. $M_{2\_attributes} = 0.78$ average distance from the average rating across all participants; $t(214) = 2.27$, $p = .024$, $d = 0.31$).

Conclusion. Across Studies 2A and 2B, we demonstrate two different mechanisms – rating repetition and the number of considered attributes – that help explain why reviewing experts have less extreme ratings. These findings provide support for a causal relationship between reviewing expertise and restrained summary ratings. Further, results from the Qunar field data (Study 1), demonstrate the generalizability of the phenomenon in the real-world.

Although we have provided considerable support for the restraint-of-expertise phenomenon, questions remain: (i) Does the restraint-of-expertise effect generalize to other real-world review platforms (not just Chinese-based but also Western-based review platforms) and to other industries (restaurants as well as hotels)? (ii) So far, the restraint-of-expertise effect has only been observed in assigned star ratings; is the effect also displayed in what reviewers write about, that this, the sentiment of the review text, even when the assigned star rating is the same? (iii) Does the attenuated impact of ratings by reviewing experts (vs. novices) on the aggregate valence metric demonstrated in Study 1 replicate on other review platforms? (iv) Which type of reviewer, experts or novices, actually benefit service providers and when does this happen? These are some of the questions that will be addressed in the following studies.

Study 3: TripAdvisor (Field Data)

Purpose. In Study 3, we test whether the restraint-of-expertise effect, H1, as observed in reviews from the Chinese-based review platform Qunar.com, (i) replicates in a North American-based review platform, TripAdvisor.com, (ii) occurs not only between reviewers (reviewing experts vs. novices), but also within reviewers (experts vs. pre-experts), and (iii) is also exhibited in the sentiment of written reviews. Further, we test two of the downstream consequences of the restraint-of-expertise effect: (iii) the impact of ratings on aggregate metrics, H3, and (iv) relative ratings between experts and novices, H4.

In this study, we collected and analyzed two sets of review data: (i) all the reviews written about a number of service providers (the *by-service-provider data*) and (ii) all the reviews generated by a number of reviewers (the *by-reviewer data*). The value of the by-service-provider data is that it allows us to address and mitigate the concern that the observed restraint-of-expertise effect may simply be an artifact of experts being less selective with the businesses they review. With the by-service-provider data, we examine the assigned star ratings *when both reviewing experts and novices select and review the same service provider*. We go on to conduct a more granular test of our restraint-of-expertise hypothesis by investigating the sentiment of the review text *when reviewing experts and novices assign the same rating for the same service provider*.

The value of the by-reviewer data is that it allows us to address the concern that the restraint-of-expertise effect may just be an artifact of novices being more selective with writing reviews, doing so only when experiences are extreme. With the by-reviewer data, we examine *how assigned ratings and sentiment of written reviews change over time for one reviewer*, across all reviewers.

Variables and Analyses. We scrape and analyze two sets of review data from TripAdvisor, a major online English-based travel review platform. The first set of data, by service provider, contains over 39,000 online reviews that were posted over a one-year time span, from 60 hotels across 6 major cities (see **Table 1** for description of dataset; see **Table 2** for variable list; see **Table 3** for summary statistics of variables). The second set of data, by reviewer, contains all the reviews (over 75,000) that were generated by 657 high contributing reviewers on TripAdvisor (we collected reviews from a number of reviewers who have generated at least 30 reviews on the TripAdvisor platform at time of data scraping).

The main independent variable of interest is *reviewing expertise*. We operationalize reviewing expertise based on TripAdvisor's platform-defined *0-6 Contributor Level*. Similar to Qunar, TripAdvisor measures their reviewing experts using a point-based system, where points are predominantly assigned for the number of reviews generated by a reviewer. We used the natural logarithm of TripAdvisor's *0-6 Contributor Level*, i.e., ln(*Contributor_level* + 1), in our analysis to normalize its distribution. Throughout the analyses, we provide descriptive statistics for the first two Contributor Levels, levels 0 and 1, and the last two Contributor Levels, levels 5 and 6. The data on reviewer's *Contributor Level* is contained in our by-service-provider data.

Similar to Study 1, the main dependent variables of interest are rating polarity and the impact of ratings on the aggregate valence metric. (For descriptions of these variables, see Study 1). We also compare the relative assigned ratings between experts and novices. We were also able to conduct text analyses to uncover (i) the polarity of the written review sentiment and (ii) the number of domain-specific (hotel) attributes in each review. Review sentiment was calculated by using two major word-sentiment dictionaries: Bing-Liu (Liu 2012) and AFINN (Hansen et al. 2011). (We used two word-sentiment dictionaries for measurement robustness of the *review sentiment* variable.) Each word in a review is associated with a specific sentiment score, based on the word-sentiment dictionary used (a score of 0 is assigned if the word is not contained in the word-sentiment dictionary). The review sentiment score is calculated by adding the sentiment value of all words in the review divided by the total number of words in the review. The *polarity of review sentiment* is calculated by taking the absolute value of the review sentiment score.

The *number of domain-specific attributes considered* was calculated using Part-of-Speech (POS) tagging (Hornik 2016). After POS tagging each word in all hotel reviews in our dataset,

we only kept the nouns. Next, we removed city-specific terms by conducting term frequency-inverse document frequency (*tf-idf*) analysis across the six cities; we also removed the term *hotel* from the list. This allowed us to compile 50 of the most frequently used hotel-related nouns; e.g., *room, lobby,* and *location* (for the full list of the 50 hotel-related nouns, see **Web Appendix C**). Next, for each review, using a match and count based algorithm, we identified the number of unique nouns mentioned in the review that were contained in the list of 50 hotel-related nouns. This produced our number of hotel-specific attributes mentioned in each review.

A key moderating variable we test is *general level of service* by the business, which is operationalized in this study by a moving user rating average, based on the most recent 20 reviews about the service provider at the time of review posting. This moderating variable is used to test H4 on relative ratings between reviewing experts and novices.

Results: *(i) Platform-Defined Reviewing Expert*. We find that reviewers operationalized as experts using TripAdvisor's *0-6 Contributor Level* do indeed exhibit greater quality-based features of expertise from our Alba and Hutchinson-based conceptual definition, in terms of (i) having a higher degree of elaboration in their reviews (by number of characters: $M_{Levels\_0\_1} = 431$ vs. $M_{Levels\_5\_6} = 740$, $r = .34$, $p < .001$; by number of words: $M_{Levels\_0\_1} = 72$ vs. $M_{Levels\_5\_6} = 110$, $r = .34$, $p < .001$), (ii) including a greater number of category-related attributes in their reviews ($M_{Levels\_0\_1} = 5.7$ vs. $M_{Levels\_5\_6} = 7.7$ hotel-related attributes considered in review, $r = .30$, $p < .001$), and (iii) having generated reviews that are deemed generally more favorable by readers ($M_{Levels\_0\_1} = 0.40$ vs. $M_{Levels\_5\_6} = 0.47$ average 'Helpful' votes per review post, $r = .07$, $p < .001$).

(ii) Expertise and Rating Evaluations (H1). We first analyze our by-service-provider data, that is, looking at all reviews for one hotel, across all our hotels. We test whether expertise in

generating reviews affects rating evaluations. Results from our mixed-effects regression model show that reviewers higher on TripAdvisor's *0-6 Contributor Levels* demonstrate greater restraint from extremes in their assigned ratings ($M_{Level\_0\_1}$ = 1.59 vs. $M_{Level\_5\_6}$ = 1.33 average distance away from midpoint of the five-point rating scale; β = -0.13, t(39135) = -28.95, p < .001, *Cohens* f^2 = 0.146; see **Figure 1B**). Results are robust when analyzing only reviews that were generated by reviewers who have generated at least 10 reviews (β = -0.25, t(19740) = -11.77, p < .001, f^2 = 0.084) and 20 reviews (β = -0.32, t(14219) = -10.09, p < .001, f^2 = 0.085), suggesting that the observed restraint-of-expertise effect is not purely driven by reviewers who have just written a few reviews.

We relax our parametric assumption about the rating polarity dependent variable by conducting a more conservative test, ordered logistic regression (Ripley et al. 2013). The analysis demonstrates robustness in the restraint-of-expertise effect (β = -0.49, t = -30.08, p < .001). As a robustness of measurement of the dependent variable, *rating polarity*, we compare the dispersion of ratings by experts and novices. Results from Bartlett's test of homogeneity of variances show that the variance of ratings by reviewing experts ($SD_{Level\_5\_6}$ = 0.85) is significantly lower than the variance of ratings by novices ($SD_{Level\_0\_1}$ = 1.02; K^2 = 367.74, p < .001).

We repeat our main analysis for H1 with our by-reviewer data, that is, looking at all reviews for one reviewer, across all reviewers, and results are robust; as reviewers generate more reviews, the ratings in their reviews become more restrained from the extremes (by rating polarity: β = -0.046, t(74928) = -6.69, p < .001, f^2 = 0.024; by variance of ratings: K^2 = 98.84, p < .001).

Next, we conduct text analyses to test the restraint-of-expertise effect on the sentiment of the review text, *while controlling for the actual assigned ratings by the reviewer*. In other words, even when reviewers assign the same ratings for the same service provider, do novice and expert reviewers use different affective language in their reviews? Our results from analyzing the by-service-provider data show that for a given hotel with the same assigned rating, reviewing experts (vs. novices) demonstrate more restraint in the polarity of the sentiment of their review text (by Bing-Liu's word-sentiment dictionary: $M_{Level\_0\_1} = 0.092$ vs. $M_{Level\_5\_6} = 0.078$, $\beta = -0.009$, $t = -23.52$, $p < .001$, $f^2 = 0.119$; by AFINN word-sentiment dictionary: $M_{Level\_0\_1} = 0.197$ vs. $M_{Level\_5\_6} = 0.159$, $\beta = -0.025$, $t = -27.36$, $p < .001$, $f^2 = 0.138$; see **Figure 1B**). Further, our results are robust when analyzing our by-reviewer data (by Bing-Liu's word-sentiment dictionary: $\beta = -0.0014$, $t = -2.99$, $p = .003$, $f^2 = 0.011$; by AFINN word-sentiment dictionary: $\beta = -0.0050$, $t = -4.77$, $p < .017$, $f^2 = 0.024$), suggesting that as reviewers gain more experience generating reviews, *even when assigning the same star rating*, the polarity of their review sentiment becomes more restrained.

(iii) Impact of expertise on shifting the aggregate valence metric (H3). Next, we test the impact of expertise on aggregate valence metrics. Consistent with Study 1 results, we find that reviewing expert (vs. novice) ratings have significantly less impact on the aggregate valence metric – both in terms of the moving valence metric ($M_{Level\_0\_1} = 0.67$ vs. $M_{Level\_5\_6} = 0.60$; $\beta = -0.06$, $t(39115) = -13.96$, $p < .001$, $f^2 = 0.071$) and the cumulative valence metric ($M_{Level\_0\_1} = 0.73$ vs. $M_{Level\_5\_6} = 0.62$; $\beta = -0.07$, $t(39136) = -17{,}74$, $p < .001$, $f^2 = 0.090$).

(v) Relative ratings between experts and novices (H4). Lastly, we test who – reviewing experts or novices – assign higher ratings, and how this might depend on the general level of service provided by the business. Using a mixed-effect regression model, where the hotel is

treated as a random effect, we test the interaction between the general level of service and TripAdvisor's measure of reviewing expertise on assigned ratings. Consistent with our theorizing, H4, we find a significant negative interaction (β = -0.11, t(39113) = -7.34, $p < .001$, f^2 = 0.037; see **Figure 3A**). To detect specific values along the general level of service where reviewing experts (vs. novices) assign systematically higher and lower ratings, we conduct a follow-up floodlight analysis (Johnson and Neymar 1936; Spiller et al. 2013). Our floodlight analysis demonstrates that for service providers that generally provide *mediocre to poor* experiences (specifically, recent average ratings below 3.8, see **Figure 3A**), reviewing experts assign significantly higher ratings ($M_{Level\_5\_6}$ = 3.55) than novices ($M_{Level\_0\_1}$ = 3.41; β = 0.09, t(2995) = 2.69, p = .007, f^2 = 0.049). For service providers that generally provide *excellent* experiences (specifically recent average ratings above 4.1), reviewing experts assign significantly lower ratings ($M_{Level\_5\_6}$ = 4.40) than novices ($M_{Level\_0\_1}$ = 4.54; β = -0.07, t(30224) = -10.48, $p < .001$, f^2 = 0.060).

* * * FIGURE 3 ABOUT HERE * * *

Conclusion. Using hotel reviews from TripAdvisor, we replicate the restraint-of-expertise effect, demonstrated not only between reviewers (reviewing experts vs. novices), but also within reviewers (expert vs. pre-expert) and evidenced not only in the assigned ratings, but also in the sentiment of the review text. Further, we demonstrate two major consequences of the restraint-of-expertise effect. First, reviewing experts (vs. novices) have less impact on the aggregate valence metric. Second, we demonstrate that reviewing experts (vs. novices) systematically benefit and harm service providers with their ratings. Specifically, for service providers that generally provide mediocre (excellent) experiences, experts assign significantly higher (lower) ratings than novices.

Although we have provided considerable support for the restraint-of-expertise hypothesis, we have only tested reviews on hotels, and the two platforms that were studied adopt a predominantly quantity-based approach to designating expertise. In Study 4, we assess whether our key findings generalize to the restaurant service domain and replicate on a more quality-based expertise designated platform.

Study 4: Yelp (Field Data)

Purpose. The purpose of Study 4 is to replicate the restraint-of-expertise effect (H1 and H2) and its downstream consequences (H4) on a different review platform: Yelp.com. The Yelp review platform is unique from the two previously studied review platforms – Qunar and TripAdvisor – in terms of (i) its adoption of an expertise designation that is based more on the quality (vs. quantity) of reviews, and (ii) its designation of expertise as binary, instead of gradient/levels. Further, in Study 4, we collect and analyze reviews about restaurants, rather than hotels, as studied in the previous two field studies, allowing us to generalize our results across service domains.

As in Study 3, we collected and analyzed two sets of review data: (i) reviews written about a number of service providers (the *by-service-provider data*) and (ii) reviews generated by a number of reviewers (the *by-reviewer data*). These two sets of data allow us to address alterative explanations related to (i) reviewers' selection of service providers and (ii) reviewers' self-selection for writing reviews, respectively.

Variables and Analyses. We collected and analyzed two sets of review data from Yelp.com, a major online restaurant review platform based in North America. The first set, *by service provider*, contains over one million online reviews from 2039 restaurants across four major cities (see **Table 1** for description of dataset; see **Table 2** for variable list; see **Table 3** for summary

statistics of variables). The second set, *by reviewer*, contains over one million reviews that were generated by 13,280 expert reviewers (i.e., reviewers that were designated as 'Elite' from Yelp at time of data collection).

The main independent variable is *reviewing expertise*. We operationalize reviewing expertise based on Yelp's platform-defined 'Elite' status designation. As stated on their Yelp's website, "Elite-worthiness is based on a number of things, including well-written reviews, high quality tips, a detailed personal profile, an active voting and complimenting record, and a history of playing well with others" (Yelp Support Center 2019). However, unlike TripAdvisor and Qunar, the designation of expertise is done by humans, where other fellow reviewers on the platform nominate a reviewer for their 'Elite' worthiness, and then a 'Community Manager' decides whether or not an official 'Elite' badge is assigned to that reviewer for the year.

Note that the Yelp data contains not only the current reviewing expertise designation ('Elite' vs. non-'Elite') at time of data collection, but also the list of all the previous years a reviewer had obtained the 'Elite' badge. This information allows us to conduct our *within reviewer* analyses, where we compare and contrast reviews generated before and after the year a reviewer obtained her first 'Elite' badge.

The main dependent variables of interest are rating polarity and assigned star ratings. We also conduct text analyses to obtain text-related measures: sentiment of review text (Liu 2012) and number of domain-specific (i.e., restaurant-related) attributes mentioned in the reviews (see **Web Appendix C**). (All of these variables were discussed in the previous field studies.)

A key moderating variable we test is *general level of service* by the business, which we operationalize using Yelp's overall star rating designation of the business, in increments of 0.5, at the time reviews were collected.

Because of the nested nature of reviews by reviewers and by restaurants, we conduct mixed-effects regression analyses. Included in the analyses are a number of control variables, including restaurant identification (as a random effect, when analyzing the by-service-provider data), reviewer identification (as a random effect, when analyzing the by-reviewer data), and date of review post (converted to number of months from date of review scraping).

Results: *(i) Platform-Defined Reviewing Expert*. Comparing between reviewers (reviewing experts vs. novices) in our *by-service-provider* data, we find that Yelp 'Elite' (vs. Yelp non-'Elite') reviewers demonstrate greater quality-based features of reviewing expertise, in terms of (i) having a higher degree of elaboration in their reviews (by characters per review: M_{Elite} = 985 vs. $M_{Non\text{-}elite}$ = 538, r = .34, p < .001; by words per review: M_{Elite} = 186 vs. $M_{Non\text{-}elite}$ = 102, r = .34, p < .001), (ii) demonstrating greater category (restaurant) knowledge in their reviews (M_{Elite} = 9.8 vs. $M_{Non\text{-}elite}$ = 6.6 restaurant attributes mentioned per review, r = .25, p < .001), and (iii) having generated reviews that are deemed more favorable by readers (M_{Elite} = 2.9 vs. $M_{Non\text{-}elite}$ = 0.7 average 'Useful' votes per review post, r = .39, p < .001).

With our *by-reviewer* data, we compare and contrast reviews that were generated prior to, versus after, 'Elite' badge designation. In line with our between reviewer results above, we find that reviews generated after (vs. before) receiving one's 'Elite' designation show greater degrees of expertise, in terms of greater degree of elaboration in the reviews (by characters per review: M_{Elite} = 903 vs. $M_{Pre\text{-}elite}$ = 695, r = .16, p < .001; by words per review: M_{Elite} = 172 vs. $M_{Pre\text{-}elite}$ = 132, r = .16, p < .001), greater degree of category knowledge (M_{Elite} = 8.0 vs. $M_{Pre\text{-}elite}$ = 6.9 restaurant attributes mentioned per review, r = .09, p < .001), and greater degree of favorability by readers (M_{Elite} = 3.0 vs. $M_{Pre\text{-}elite}$ = 1.7 average 'Useful' votes per review post, r = .14, p <

.001. This is a conservative estimate, as it does not account for the fact that reviews generated prior to (vs. after) 'Elite' designation has been available online for a longer period of time.

(ii) Expertise and Rating Evaluations (H1). Consistent with results from the previous field studies and experiments, we find evidence for the restraint-of-expertise hypothesis between expert and novice reviewers when comparing by rating polarity (M_{Elite} = 1.12 vs. $M_{Non\text{-}elite}$ = 1.43 average distance from midpoint of 5-point scale; β = -0.28, t(1019938) = -164.61, $p < .001$, *Cohen's* $f^2 = 0.163$; see **Figure 1C**) as well as by variance in assigned ratings (SD_{Elite} = 1.01 vs. $SD_{Non\text{-}elite}$ = 1.34; K^2 = 24,111, $p < .001$). More importantly, we observe the restraint-of-expertise effect *within* expert reviewers (by rating polarity: M_{Elite} = 1.10 vs. $M_{Pre\text{-}elite}$ = 1.19; β = -0.026, t(1008538) = -12.04, $p < .001$, $f^2 = 0.012$; and by variance in ratings: SD_{Elite} = 1.06 vs. $SD_{Pre\text{-}elite}$ = 1.17; K^2 = 2,606, $p < .001$).

Next, we test whether expert, versus novice, reviewers express more restraint in the sentiment of their review text when controlling for the assigned ratings by the reviewers. Consistent with the TripAdvisor results, we find that *even when expert and novice reviewers assign the same ratings for the same service provider*, expert (vs. novice) reviewers demonstrate more restraint in the polarity of the sentiment of their review text (by Bing-Liu's dictionary: M_{Elite} = 0.048 vs. $M_{Non\text{-}elite}$ = 0.066, β = -0.016, t = -13.68, $p < .001$, $f^2 = 0.014$; by AFINN dictionary: M_{Elite} = 0.163 vs. $M_{Non\text{-}elite}$ = 0.111, β = -0.048, t = -16.18, $p < .001$, $f^2 = 0.016$; see **Figure 1C**). These results are robust when comparing pre- vs. post Yelp 'Elite' status designation (by Bing-Liu's dictionary: M_{Elite} = 0.047 vs. $M_{Pre\text{-}elite}$ = 0.056, β = -0.008, t = -8.96, $p < .001$, $f^2 = 0.009$; by AFINN dictionary: M_{Elite} = 0.104 vs. $M_{Pre\text{-}elite}$ = 0.128, β = -0.021, t = -9.70, $p < .001$, $f^2 = 0.010$).

(iii) Mechanism: Attributes Considered (H2). Regarding H2, we test whether the number of considered attributes explains the restraint-of-expertise effect. As a measure of the number of considered attributes, we use the number of domain-specific (restaurant-related) nouns mentioned in the reviews, which was extracted using Part-of-Speech tagging (see Study 3 for details on this process). Using mediation analysis in R (*mediation* R package, Tingley et al. 2017), we find that number of considered attributes mediates the effect of reviewer expertise on restraint ratings in both our *between* reviewers (-0.0470, 95% CI: -0.0477 to -0.0463, 1000 iterations, 15.2% proportion of main effect mediation) and our *within* reviewers analyses (-0.0153, 95% CI: -0.0156 to -0.0150, 1000 iterations, 15.7% proportion of main effect mediated).

(iv) Relative ratings between experts and novices (H4). Lastly, we test who – reviewing experts or novices – assign higher ratings, and how this might depend on the general level of service provided by the business. Results from our mixed-effects regression model show that there is a significant negative interaction between the general level of service and Yelp's expert reviewer on relative assigned ratings (β = -0.24, t = -40.23, $p < .001$, $f^2 = 0.040$; see **Figure 3B**).

Specifically, we see that for restaurants with 2.0, 2.5, 3.0, and 3.5 average star ratings, experts, on average, assigned significantly *higher* ratings than novices by 0.41, 0.34, 0.24, and 0.12, respectively (all p's < .001). In contrast, for restaurants with 4.5 and 5.0 average star ratings, reviewing experts assigned significantly *lower* ratings than novices by 0.12 and 0.07, respectively (both p's < .001).

Conclusion. Using restaurant reviews from Yelp, we demonstrate the restraint-of-expertise effect (H1), using both assigned ratings and review sentiment. We demonstrate this both between reviewers (experts vs. novices) and within reviewers (experts vs. pre-experts). We provide evidence for the mechanism of number of attributes considered (H2). Finally, we replicate a

major consequence of the restraint-of-expertise effect. Expert (vs. novice) reviewers systematically benefit and harm service providers with their ratings, depending on the general level of service of the business (H4).

General Discussion

In this research, we study reviewing experts on online review platforms. Our main hypothesis is that greater expertise in generating reviews leads to greater restraint from extremes summary evaluations. Across five studies (three field studies and two experiments), we test this restraint-of-expertise hypothesis, its explanation, and its consequences for service providers, such as hotels and restaurants. The restraint-of-expertise hypothesis is observed across three different review platforms (TripAdvisor, Qunar, and Yelp), is shown using both assigned ratings and review text sentiment, and demonstrated not only between reviewers (experts vs. novices), but also within reviewers (expert vs. pre-expert), mitigating concerns related to reviewers' selection of service providers, and reviewer's self-selection in writing reviews. Two experiments replicate the main effect and provide support for an attributes-based explanation. The field studies demonstrate two major consequences of the restraint-of-expertise effect. (i) Reviewing experts (vs. novices) play a lesser role in shifting the aggregate valence metric over time. (ii) Reviewing experts systematically benefit and harm service providers with their ratings. For service providers that generally provide mediocre (excellent) experiences, experts assign significantly higher (lower) ratings than novices.

There are three important theoretical implications of our work. First, our research extends the literature on expertise to the online user-generated content (UGC) domain. Much of the extant research on expertise was conducted in predominantly offline domains, such as playing chess (Charness et al. 2005; Gobet and Simon 1998), solving physics problems (Chi, Feltovich,

and Glaser 1981; Larkin et al. 1980), and tasting wines (Latour and Dayton 2018; Solomon 1990). However, given the rise of UGC and the ability of UGC platforms to identify top users, it has been unclear whether much of what we already know in the expertise literature can be applied to the online UGC domain. Admittedly, various aspects about UGC platforms are novel, such as their extremely large-scale nature and their lack of formal qualifying tests to designate expertise levels. Our research demonstrates that these so-called online 'expert' users, by and large, display features of expertise, as defined in the prior literature (Alba and Hutchinson 1987). We acknowledge the imperfection in capturing expertise with quick and scalable approaches, such as a point-base system, especially one that places greater weight on quantity over quality. However, we concede that such approaches are practically reasonable, given the large-scale nature of many UGC platforms. Future research can help refine efficient scalable approaches that more effectively capture expertise.

Second, our research contributes to the discussion on the observed extreme rating distribution phenomenon in online reviews. Much of the prior attribution for this observed pattern has been the reviewer's motivation to generate reviews (a self-selection bias; Hu, Pavlou, and Zhang 2009). We agree that self-selection plays an important role in affecting the extent of observed extreme rating distributions, however, our research points to another important factor, *reviewing expertise*. Novice reviewers evaluate in a more polarizing manner (Linville 1982; Rozin, Ashmore and Markwith 1996), but as they gain greater experience generating reviews, they (implicitly) consider more attributes in their evaluations, and in turn, provide summary ratings that are more restrained from extremes. Hence, the overall degree of expertise of reviewers for a particular product/service influences the extent to which extreme rating distributions are observed.

Third, our research contributes to the literature concerning the (counter-) influential nature of experts on consumer choice (Biswas et al. 2006; Packard and Berger 2017). For example, Biswas et al. (2006) find that the influential nature of expert endorsers (compared to celebrity endorsers), in terms of reducing perceived risk, is particularly pronounced for high technology-oriented products (e.g., computers, high-definition televisions) versus low technology-oriented products (e.g., treadmills, mattresses). Packard and Berger (2017) show that novices are more likely to use explicit endorsement styles in their reviews (e.g., "I recommend it" vs. "I like it"), which are found to be more persuasive and increase purchase intent. These researchers suggest that *ceteris paribus*, the endorsement styles novices and experts tend to use can lead to greater persuasion by novices. In our research, we demonstrate how the restraint-of-expertise effect also dampens the influential nature of reviewing experts. Because experts generally assign ratings that are less polarizing, and user rating averages by their nature are skewed from extremes, we find that expert reviewers (vs. novices) have an attenuated impact on shifting the cumulative and moving user rating averages (note that this does not yet account for the fact that there are substantially more novices than experts, about 19 to 1 in our Yelp data). Aggregate metrics, such as the user rating average, are important as they are known to affect service-provider page rank (Ghose, Ipeirotis, and Li 2012) and consumer consideration (Luca 2016; Vermeulen and Daphne 2009). So, although the actual review content generated by reviewing experts may be more favored by consumers (Racherla and Friske 2012; Zhang, Zhang, and Yang 2016), *in the context where information is abundant and aggregated*, such as the case with online reviews, the attenuated impact experts have on the aggregate valence metric means that reviewing experts (vs. novices) play a lessor role on the service providers consumers consider before reading individual reviews.

We acknowledge that there are different approaches to calculating the aggregate valence metric; some platforms use a simple arithmetic average, others use algorithms that place greater weight on the number of past reviews generated by the reviewer, the number of 'Like' votes received by the review, whether or not the review is a verified purchase, and/or the recency of the review (Matsakis 2019). Some of these approaches can certainly mitigate the reduced impact reviewing experts have on the aggregate valence metric. The reweighting threshold that is required to offset the attenuated impact of reviewing experts remains unclear, because of the restrained ratings of reviewing expert and the outnumbering of novices over experts. Hence, future research can explore optimal weighting strategies to mitigate this concern (e.g., Dai et al. 2018).

Our research has three important practical implications for businesses. First, our research provides caveats to the common practice of companies actively seeking and incentivizing reviewing experts. We acknowledge and find that reviews by experts (vs. novices) do, on average, receive more favorability (e.g., 'Like') votes by readers. However, we argue that review favorability is only one aspect that shapes consumer choice. Another important aspect is the aggregate valence metric. Much of the research on online reviews has emphasized the importance of the aggregate valence metric (Babić Rosario et al. 2016; Floyd et al. 2014; You, Vadakkepath, and Joshi 2015), as it affects the page on which service providers appear in searches (Ghose, Ipeirotis, and Li 2012) and whether or not consumers consider the service provider as an option (Fisher, Newman, and Dhar 2018; Vermeulen and Seegers 2009). In other words, a reviewer can provide a highly detailed account of a restaurant experience, but if that restaurant is not even considered by readers, the impact of the review is largely attenuated. Given the importance of aggregate metrics, who tends to elevate/lower a service providers' aggregate

valence metric and when? We find that it depends on the general level of service by the service provider. Service providers that generally provide *excellent* levels of service should be cautious offering reviewing experts incentives to review, as experts are hesitant to give out 5-star ratings. Because of their more polarizing rating approach, novices (vs. experts) are more likely to assign 5-star ratings for positive experiences, and hence elevate the service provider's valence metric. In contrast, service providers that generally provide *mediocre* service can greatly benefit from reviews by experts, as experts, for such service providers, assign consistently higher ratings than their novice counterparts. It is important to note that these recommendations are based on elevating a service providers' aggregate valence metric. We acknowledge that if service providers are seeking consumer reviews as feedback to help improve the business, as opposed to using consumer reviews as part of the marketing mix, the highly detailed accounts by reviewing experts are highly valuable, regardless of the service level.

Second, our research brings to light the issue of adopting rating scales with the same granularity for experts and novices, and the problem associated with combining expert and novice ratings to form aggregate valence metrics. Across three different review platforms, we observe that reviewing experts (vs. novices) are more likely to assign ratings that are less extreme; their rating distribution is akin to an inverse U-shaped (vs. J-shaped; see **Figure 1C**). Further, when comparing the averages of expert ratings to those of novices, we find that they are *not* the same; this finding is in line with past research showing differences between expert judgment and lay people's opinions (de Langhe, Fernbach, and Lichtenstein 2016; Holbrook 1999). Therefore, we recommend review platforms adopt different rating scales for their expert and novice users (using a more granular scale for their experts) and present different aggregate valence metrics for ratings by these two groups. One can see this approach with platforms such

as Rotten Tomatoes, where critics evaluate on a 10-point scale and the audience evaluates on a 5-point scale, and the aggregate scores for critics and the audience are separated. There are several caveats to consider with these recommendations. First, review platforms that designate expertise along multiple levels (e.g., Qunar's 1-7 Expertise Levels) would need to consider a cut-off point(s) in order to assign users scales with the appropriate granularity. Second, an important gap that remains to be addressed is whether review-reading consumers would rely more on novice or expert aggregate ratings and when. This is an important concern that remains to be addressed in future research.

Third, our research provides review platforms with a strategy to reduce the degree of user rating extremity. We recommend platforms have their users evaluate along the product/service attributes before assigning a summary rating. Our research shows that considering many (vs. few) attributes of a product/service experience, prior to assigning the summary rating, reduces the extremity of the summary rating (see Study 2B). However, an important caveat to consider is that having users consider too many attributes will reduce the likelihood users complete the review, hence, lowering the review count. Future research can investigate the optimal number of attributes that (i) reduces rating extremity but also (ii) minimizes hindrance of review completion.

The focus of our research is on the relationship between reviewing expertise and rating evaluations. Although our analyses include some measures of consumer perceptions of reviews (e.g., 'Like', 'Helpful', and 'Useful' votes by readers), the relationship between the review-reading consumers and expert-generated reviews remains an important area for future research. A number of questions remain to be answered: How do review-reading consumers perceive review content generated by reviewing experts? What role does the expertise badge (e.g., 'Elite

2020') have on how readers perceive an expert-generated review, if any? Are there specific circumstances where the expertise badge does and does not matter? If so, what are these circumstances? Overall, how might the findings on the relationship between reader and expert-generated review shape the choices review platforms make in designing their platform interface? We believe these are some important questions that remain to be answered in the area of reviewer expertise.

To conclude, we provide evidence, in the context of consumer-generated reviews, of how reviewing expertise affects rating evaluations, and the downstream consequences of expert ratings for businesses. The findings are important to service providers and rating platforms, particularly as consumers move away from traditional offline media and towards online digital media, where user-generated content plays an increasingly larger role in shaping consumer choice.

Table 1. Description of the Qunar, TripAdvisor, and Yelp Datasets

	Qunar (Study 1)	**TripAdvisor (Study 3)**		**Yelp (Study 4)**	
Reviews collected based on	Service Provider	Service Provider	Reviewer	Service Provider	Reviewer
Language	Chinese	English	English	English	English
Number of Cities	4	6	NA	4	NA
List of Cities	Beijing, Gaungzhou, Sanya, Shanghai	Chicago, HK, London, Los Angeles, Paris, Singapore	NA	Las Vegas, Phoenix, Pittsburgh Toronto,	NA
Service Provider Type	Hotel	Hotel	Predominantly Hotel	Restaurant	Predominantly Restaurant
Total Number of Service Providers	60	60	NA	2039	NA
Number of reviewers	NA	NA	657	NA	13,280
Number of reviews	125,985	39,203	75,587	1,021,978	1,021,819
Date of Data Collection	March 2016	January 2017	October 2019	January 2018	January 2018
Date of Reviews	2007 – 2016	02-2016 – 01-2017	2009 – 2019	2004-2018	2004-2018

Notes:
Qunar & TripAdvisor:
Reviews from Qunar and TripAdvisor were scrapped from their online website: https://www.qunar.com/ and https://www.tripadvisor.ca/
For the by-service provider data from Qunar and TripAdvisor, selection of hotels was based on popularity on the platform at the time of data scraping. While we collected and analyzed all the review data available in the selected hotels on Qunar, we only collected and analyzed the most recent 1 year of review data on TripAdvisor.
For the by-reviewer data from TripAdvisor, we randomly selected reviewers who had posted at least 30 reviews at time of data scraping and collected all their reviews.
Yelp:
Yelp review data was compiled from the data provided by Kaggle.com: https://www.kaggle.com/yelp-dataset/yelp-dataset
For the by-service-provider data, we randomly selected restaurants located in the four cities that contained the greatest number of restaurants.
For the by-reviewer data, we randomly selected reviewers who had receive an 'Elite' badge designation, along with the reviews they had posted.

Table 2. Description of Variables

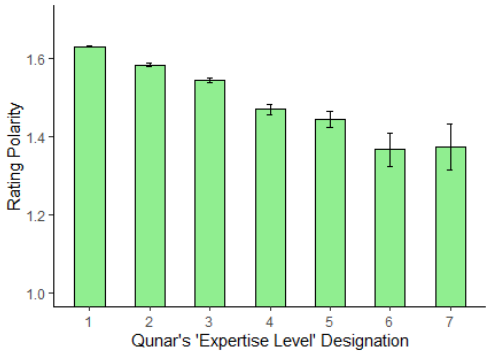
Variable	Description
Favorability	Number of favorability votes by reader (Qunar = 'Like' votes, TripAdvisor = 'Helpful' votes, Yelp = 'Useful' votes)
Length	Number of characters in the review.
MonthsAgo	Number of months ago review was posted from date of data collection.
Purpose	Categorical variable indicating purpose of the trip: family, couple, business, friends, single, unknown.
Rating	Integer star rating assigned by reviewer in the review, from 1 – *Terrible* to 5 – *Excellent.*
RatingPolarity	Distance of assigned rating from the midpoint of 3 on 5-point rating scale. Measured as the absolute value of the Rating subtracted by the scale-midpoint value of 3, i.e., |*Rating* – 3|.
ReviewerID	Identification of reviewer; only included in data that was collected by reviewers. Treated as random effect in the mixed models testing by-reviewer data
ReviewerExpertise	Platform-defined reviewer expertise (Qunar = *1-7 Expertise Level*, TripAdvisor = *0-6 Contributor Level*, Yelp = *Elite* reviewer designation.)
ServiceProvider	Identification of hotel/restaurant to which the review is attributed. Treated as random effects in the mixed models testing by-service-provider data

Table 3. Key Summary Statistics of Variables

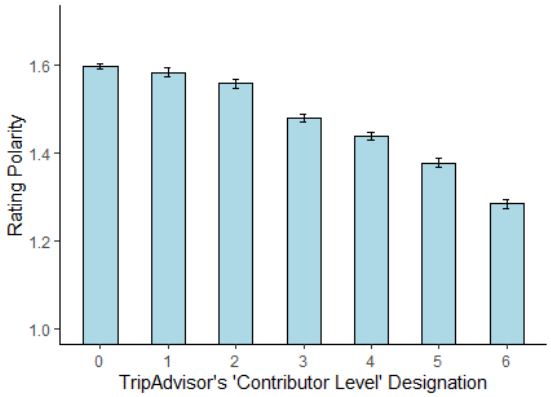
	Qunar (Study 1)				**TripAdvisor** (Study 3)								**Yelp** (Study 4)							
	By-service-provider (N = 125,985)				By-service-provider (N = 39,203)				By-reviewer (N = 75,587)				By-service-provider (N = 1,021,978)				By-reviewer (N = 1,021,819)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Favorability	0.4	2.7	0	219	0.5	0.9	0	14	0.3	0.7	0	9	1.1	3.0	0	246	2.8	5.4	0	1,608
Length	83.8	185.9	1	7,306	586.4	514.6	86	8,605	456.7	347.4	48	6955	631.5	592.4	4	9,321	867.4	659.3	1	9,321
MonthsAgo	14.0	7.9	1	101	6.9	3.2	1	12	NA	NA	NA	NA	39.4	28.8	1	150	NA	NA	NA	NA
Rating	4.46	0.91	1	5	4.33	0.95	1	5	4.1	0.94	1	5	3.85	1.28	1	5	3.77	1.08	1	5
RatingPolarity	1.61	0.62	0	2	1.49	0.67	0	2	1.27	0.70	0	2	1.36	0.71	0	2	1.12	0.72	0	2
ReviewerExpertise	1.52	0.88	1	7	2.53	2.07	0	6	NA	NA	NA	NA	0.21	0.41	0	1	NA	NA	NA	NA

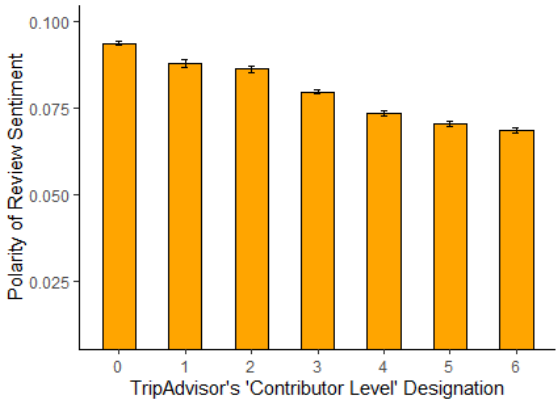
Figure 1. Polarity of Evaluations as a Function of Platform-Defined Reviewer Expertise.

A) **Qunar** (Study 1)



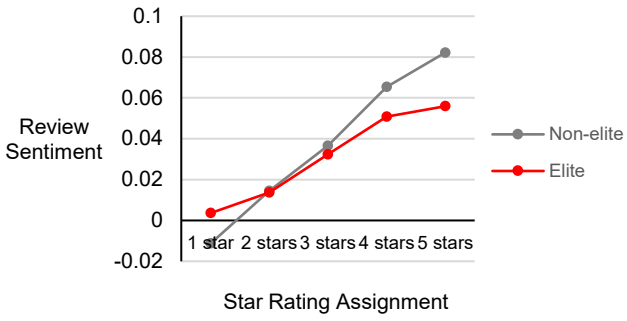
B) **TripAdvisor** (Study 3)





Review sentiment calculated using the LIU sentiment-word dictionary (Liu 2012).

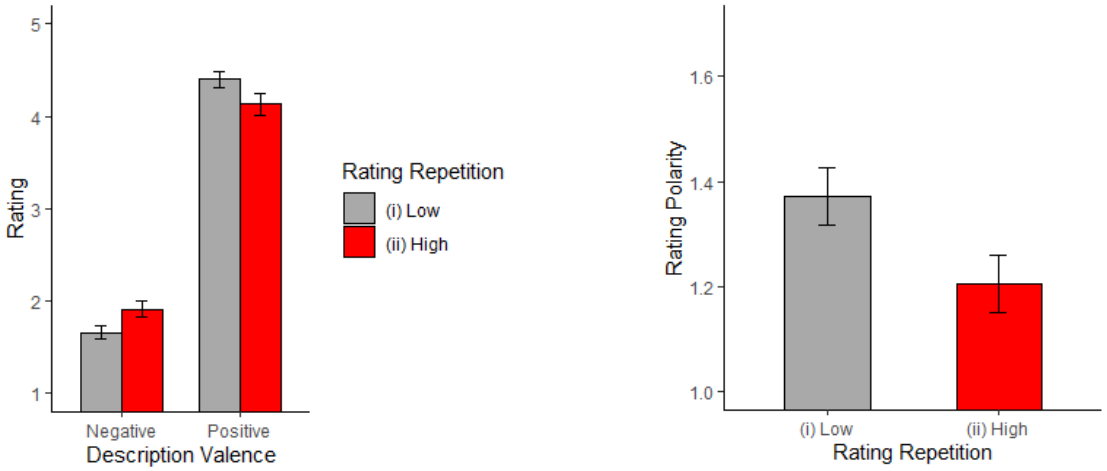
C) **Yelp** (Study 4)



Review sentiment calculated using the LIU sentiment-word dictionary (Liu 2012).

Figure 2. Study 2A and 2B Results

A) Study 2A



B) Study 2B

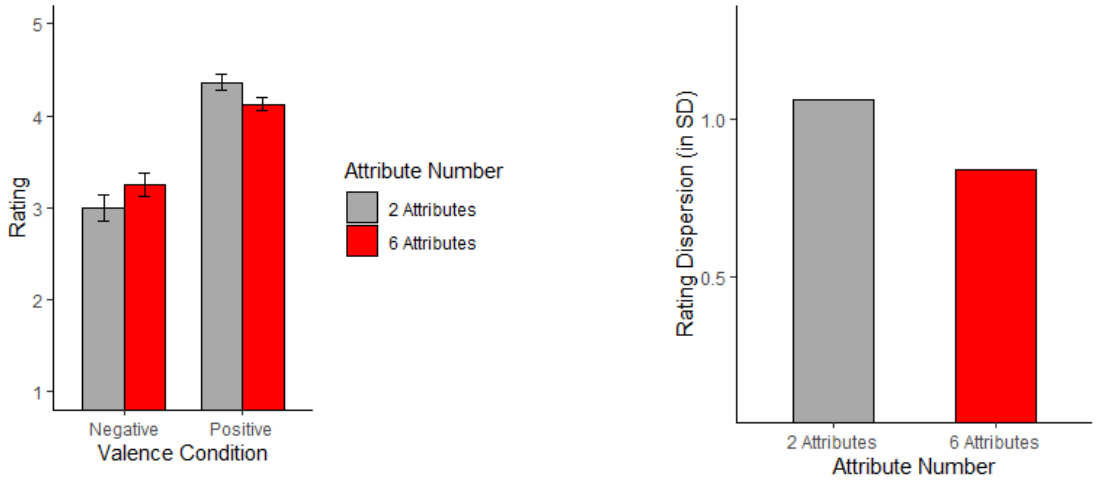
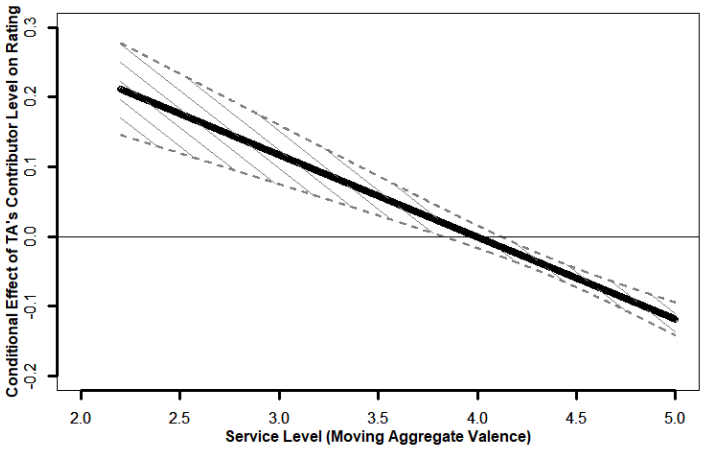
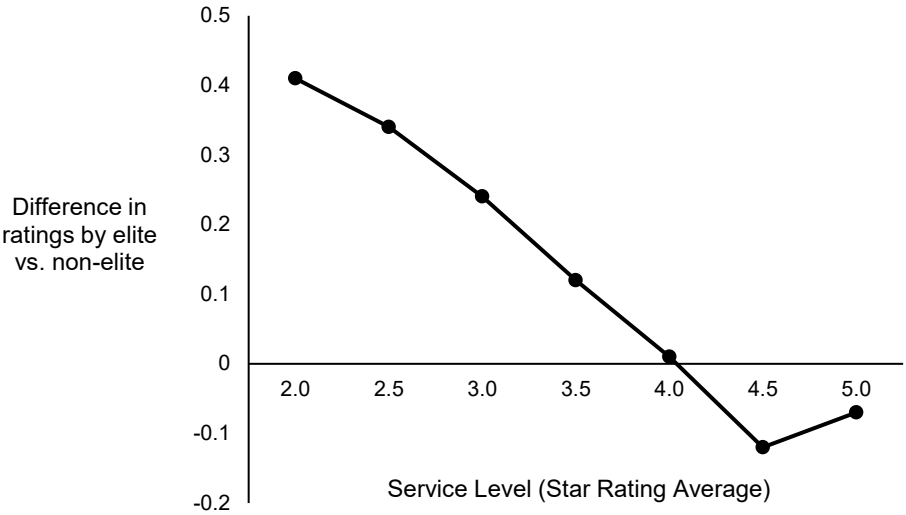


Figure 3. Difference in Ratings Between Experts and Novices as a Function General Level of Service by Service Providers.

A) TripAdvisor (Study 3)



B) Yelp (Study 4)



References

Alba, Joseph W., and J. Wesley Hutchinson (1987), "Dimensions of Consumer Expertise," *Journal of Consumer Research,* 13 (4), 411-54.

Amabile, T. M. (1983), “Brilliant but Cruel: Perceptions of Negative Evaluators,” *Journal of Experimental Social Psychology*, 19 (83), 146-56.

Babić Rosario, Ana, Francesca Sotgiu, Kristine De Valck, and Tammo HA 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.

Biswas, Dipayan, Abhijit Biswas, and Neel Das (2006), "The Differential Effects of Celebrity and Expert Endorsements on Consumer Risk Perceptions. The Role of Consumer Knowledge, Perceived Congruency, and Product Technology Orientation," *Journal of Advertising,* 35 (2), 17-31.

Chakrabarti, Meghna (2013), “Top Amazon Reviewers Get Big Perks”, *NPR*, https://www.npr.org/transcripts/247833514?storyId=247833514?storyId=247833514

Charness, Neil, Michael Tuffiash, Ralf Krampe, Eyal Reingold, and Ekaterina Vasyukova (2005), "The Role of Deliberate Practice in Chess Expertise," *Applied Cognitive Psychology*, 19 (2), 151-65.

Cheung, Christy MK, Matthew KO Lee, and Neil Rabjohn (2008), "The Impact of Electronic Word-of-Mouth: The Adoption of Online Opinions in Online Customer Communities," *Internet Research,* 18 (3), 229-47.

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.

Chi, Michelene TH, Paul J. Feltovich, and Robert Glaser (1981), "Categorization and Representation of Physics Problems by Experts and Novices," *Cognitive Science*, 5 (2), 121-52.

Chocarro, Raquel, and Mónica Cortiñas (2013), "The Impact of Expert Opinion in Consumer Perception of Wines," *International Journal of Wine Business Research*, 25 (3), 227-48.

Dai, Weijia (Daisy), Ginger Zhe Jin, Jungmin Lee, and Michael Luca (2018), "Aggregation of Consumer Ratings: An Application to Yelp.com," *Quantitative Marketing and Economics*, 16 (3), 289-339.

De Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42 (6), 817-33.

Einhorn, Hillel J. and Robin M. Hogarth (1981), "Behavioral Decision Theory: Processes of Judgment and Choice," in *Annual Review of Psychology,* Vol. 32, eds. Mark R. Rosenzweig and Lyman W. Porter, Palo Alto, CA: Annual Reviews, Inc., 53-88.

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.

Fiske, Susan T., Donald R. Kinder, and W. Michael Larter (1983), "The Novice and the Expert: Knowledge-Based Strategies in Political Cognition," *Journal of Experimental Social Psychology*, 19, 381-400.

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

Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li (2012), "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content," *Marketing Science*, 31 (3), 493-520.

Gobet, Fernand, and Herbert A. Simon (1998), "Expert Chess Memory: Revisiting the Chunking Hypothesis," *Memory*, 6 (3), 225-55.

Grewal, Lauren, and Andrew T. Stephen (2019), "In Mobile We Trust: The Effects of Mobile Versus Non-Mobile Reviews on consumer Purchase Intentions," *Journal of Marketing Research*, 56 (5), 891-808.

Hansen, Lars Kai, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter (2011), "Good Friends, Bad News-Affect and Virality in Twitter," In *Future Information Technology*, Springer, Berlin, Heidelberg.

Harmon, Robert R., and Kenneth A. Coney (1982), "The Persuasive Effects of Source Credibility in Buy and Lease Situations," *Journal of Marketing Research*, 19 (2), 255-60.

Hintzman, Douglas L. (1976), "Repetition and Memory," in *The Psychology of Learning and Motivation*, Vol. 10, ed. Gordon H. Bower, 47-91.

Holbrook, Morris B. (1999), "Popular Appeal Versus Expert Judgments of Motion Pictures," *Journal of Consumer Research*, 26 (2), 144-55.

Hong, Sung-Tai, and Robert S. Wyer Jr. (1989), "Effects of Country-of-Origin and Product-Attribute Information on Product Evaluation: An Information Processing Perspective," *Journal of Consumer Research*, 16 (2), 175-87.

Hornik, Kurt (2016). "Apache OpenNLP Tools Interface." <https://cran.r-project.org/web/packages/openNLP/openNLP.pdf>

Hoyer, Wayne D. (1984), "An Examination of Consumer Decision Making for a Common Repeat Purchase Product," *Journal of Consumer Research,* 11(3), 822-29.

Hu, Nan, Ling Liu, and Jie Jennifer Zhang (2006), "Can Online Reviews Reveal a Product's True Quality? Empirical findings and analytical modeling of online word-of-mouth communication", In *Proceedings of the 7th ACM conference on Electronic Commerce,* 324-30.

Hu, Nan, Ling Liu, and Jie Jennifer Zhang (2008), "Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects," *Information Technology and Management,* 9 (3), 201-14.

Hu, Nan, Paul A. Pavlou, and Jie Zhang (2009), "Overcoming the J-shaped Distribution of Product Reviews," *Communications of the ACM*, 52 (10), 144-7.

Johnson, Kathy E., and Carolyn B. Mervis (1997), "Effects of Varying Levels of Expertise on the Basic Level of Categorization," *Journal of Experimental Psychology: General,* 126 (3), 248-77.

Johnson, Palmer O. and Jerzy Neyman (1936), "Tests of Certain Linear Hypotheses and Their Application to Some Educational Problems," *Statistical Research Memoirs*, 1, 57–93.

Karmarkar, Uma R., and Zakary L. Tormala (2009), "Believe Me, I Have No Idea What I'm Talking About: The Effects of Source Certainty on Consumer Involvement and Persuasion," *Journal of Consumer Research*, 36 (6), 1033-49.

Korfiatis, Nikolaos, Elena García-Bariocanal, and Salvador Sánchez-Alonso (2012), "Evaluating Content Quality and Helpfulness of Online Product Reviews: The Interplay of Review Helpfulness vs. Review Content," *Electronic Commerce Research and Applications*, 11 (3), 205-17.

Larkin, Jill, John McDermott, Dorothea P. Simon, and Herbert A. Simon (1980), "Expert and Novice Performance in Solving Physics Problems," *Science*, 208 (4450), 1335-42.

LaTour, Kathryn A. and John A. Deighton (2018), "Learning to Become a Taste Expert," *Journal of Consumer Research*, forthcoming.

Linville, Patricia W. (1982), "The Complexity-Extremity Effect and Age-Based Stereotyping," *Journal of Personality and Social Psychology*, 1982 (42), 193-211.

Linville, Patricia W. and Edward E. Jones (1980), "Polarized Appraisals of Out-Group Members," *Journal of Personality and Social Psychology*, 38 (5), 689-703.

Liu, Bing (2012). "Sentiment Analysis and Opinion Mining." *Synthesis Lectures on Human Language Technologies,* 5 (1), 1-167.

Liu, Zhiwei, and Sangwon Park (2015), "What Makes a Useful Online Review? Implication for Travel Product Websites," *Tourism Management,* 47, 140-51.

Luca, M., (2016), "Reviews, Reputation, and Revenue: The Case of Yelp.com," *Harvard Business School NOM Unit Working Paper*, 12-016.

Mandler, Jean M., and Nancy S. Johnson. (1977), "Remembrance of Things Parsed: Story Structure and Recall," *Cognitive Psychology,* 9 (1), 111-51.

Matsakis, Louise (2019), "What Do Amazon's Star Ratings Really Mean?", *Wired*, https://www.wired.com/story/amazon-stars-ratings-calculated/

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.

Mollick, Ethan, and Ramana Nanda (2016), "Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts," *Management Science*, 62 (2), 1533-53.

Mishra, Debi Prasad, Jan B. Heide, and Stanton G. Cort (1998), "Information Asymmetry and Levels of Agency Relationships," *Journal of Marketing Research*, 277-95.

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.

Nowlis, Stephen M., and Itamar Simonson (1996), "The Effect of New Product Features on Brand Choice," *Journal of Marketing Research,* 33 (1), 36-46.

Packard, Grant, and Jonah Berger (2017), "How Language Shapes Word of Mouth's Impact," *Journal of Marketing Research,* 54 (4), 572-88.

Peng, Chih-Hung, Dezhi Yin, Chih-Ping Wei, and Han Zhang (2014), "How and When Review Length and Emotional Intensity Influence Review Helpfulness: Empirical Evidence from Epinions.com," *Thirty Fifth International Conference of Information Systems*, 1-16.

Racherla, Pradeep, and Wesley Friske (2012), "Perceived 'Usefulness' of Online Consumer Reviews: An Exploratory Investigation across Three Services Categories," *Electronic Commerce Research and Applications,* 11 (6), 548-59.

Rajaraman, Shiva (2009), YouTube Google Blog, https://youtube.googleblog.com/2009/09/five-stars-dominate-ratings.html

Ripley, Brian, Bill Venables, Douglas M. Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, and Maintainer Brian Ripley (2013), "Package 'mass'," *Cran R.*

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

Schoenmüller, Verena, Oded Netzer, and Florian Stahl (2019), "The Extreme Distribution of Online Reviews: Prevalence, Drivers and Implications," Working Paper.

Solomon, Gregg Eric Arn (1990), "Psychology of Novice and Expert Wine Talk," *The American Journal of Psychology,* 495-517.

Sonnier, Garrett P., Leigh McAlister, and Oliver J. Rutz (2011), "A Dynamic Model of the Effect of Online Communications on Firm Sales," *Marketing Science,* 30 (4), 702-16.

Spiller, Stephen A., Gavan J. Fitzsimons, John G. Lynch Jr, and Gary H. McClelland (2013), "Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression," *Journal of Marketing Research*, 50 (2), 277-88.

Stigler, Stephen M. (1997), "Regression Towards the Mean, Historically Considered," *Statistical Methods in Medical Research*, 6 (2), 103-14.

Stone (2014), "Elite Yelpers Hold Immense Power, and They Get Treated Like Kings by Bars and Restaurants Trying to Curry Favor", *Business Insider*, http://www.businessinsider.com/how-to-become-yelp-elite-2014-8

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai (2014), "Mediation: R package for causal mediation analysis."

Tripadvisor 2020, https://www.tripadvisor.com/TripCollectiveFAQ

Vermeulen, Ivar E., and Daphne Seegers (2009), "Tried and Tested: The Impact of Online Hotel Reviews on Consumer Consideration," *Tourism Management,* 30 (1), 123-27.

Yelp Support Center 2020, https://www.yelp-support.com/article/What-is-Yelps-Elite-Squad?l=en_US

Yin, Dezhi, Samuel D. Bond, and Han Zhang (2017), "Keep Your Cool or Let It Out: Nonlinear Effects of Expressed Arousal on Perceptions of Consumer Reviews," *Journal of Marketing Research,* 54 (3), 447-63.

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.

Zhang, Ziqiong, Zili Zhang, and Yang Yang (2016), "The Power of Expert Identity: How Website-Recognized Expert Reviews Influence Travelers' Online Rating Behavior," *Tourism Management,* 55, 15-24.