8-6-2019 1:00 PM

Two Essays on Consumer-Generated Reviews: Reviewer Expertise and Mobile Reviews

Peter Nguyen, *The University of Western Ontario*

Supervisor: Wang, Xin (Shane), *The University of Western Ontario*
: Cotte, June, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Business

Abstract

Over the past few decades, the internet has risen to prominence, enabling consumers to not only quickly access large amounts of information, but also openly share content (e.g., blogs, videos, reviews) with a substantially large number of fellow consumers. Given the vast presence of consumers in the online space, it has become increasingly critical for marketers to better understand the way consumers share, and learn from, consumer-generated content, a research area known as electronic word-of-mouth. In this dissertation, I advance our understanding about the shared content generated by consumers on online review platforms. In Essay 1, I study why and how the *expertise* of consumers in generating reviews systematically shapes their rating evaluations and the downstream consequences this has on the aggregate valence metric. I theorize, and provide empirical evidence, that greater expertise in generating reviews leads to greater restraint from extremes in evaluations, which is driven by the number of attributes considered by reviewers. Further, I demonstrate two major consequences of this restraint-of-expertise effect. (i) Expert (vs. novice) reviewers 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, experts assign significantly higher (lower) ratings than novices. Building on my investigation of expert reviewers, in Essay 2, I investigate the differential effects of generating reviews on *mobile devices* for expert and novice reviewers. I argue, based on Schema Theory, that expert and novice reviewers adopt different "strategies" in generating mobile reviews. Because of their review-writing experience, experts develop a review-writing schema, and compared to novices, place greater emphasis on the consistency of

various review aspects, including emotionality of language and attribute coverage in their mobile reviews. Accordingly, although mobile (vs. desktop) reviews are shorter for both experts and novices, I show that experts (novice) generate mobile reviews that contain a slight (large) increase in emotional language and are more (less) attribute dense. Drawing on these findings, I advance managerial strategies for review platforms and service providers, and provide avenues for future research.

Keywords: Electronic word of mouth, Expertise, Mobile devices, Online reviews, Platform strategy

Summary for Lay Audience

Over the past few decades, the internet has risen to prominence, enabling consumers to not only quickly access large amounts of information, but also openly share content (e.g., blogs, videos, reviews) with a substantially large number of fellow consumers. Given the vast presence of consumers in the online space, it has become increasingly critical for marketers to better understand the way consumers share, and learn from, consumer-generated content, a research area known as electronic word-of-mouth. In this dissertation, I advance our understanding about the shared content generated by consumers on online review platforms. In Essay 1, I study why and how the *expertise* of consumers in generating reviews systematically shapes their rating evaluations and the downstream consequences this has on the aggregate valence metric. I theorize, and provide empirical evidence, that greater expertise in generating reviews leads to greater restraint from extremes in evaluations, which is driven by the number of attributes considered by reviewers. Further, I demonstrate two major consequences of this restraint-of-expertise effect. (i) Expert (vs. novice) reviewers 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, experts assign significantly higher (lower) ratings than novices. Building on my investigation of expert reviewers, in Essay 2, I investigate the differential effects of generating reviews on *mobile devices* for expert and novice reviewers. I argue, based on Schema Theory, that expert and novice reviewers adopt different "strategies" in generating mobile reviews. Because of their review-writing experience, experts develop a review-writing schema, and compared to novices, place greater emphasis on the consistency of

various review aspects, including emotionality of language and attribute coverage in their mobile reviews. Accordingly, although mobile (vs. desktop) reviews are shorter for both experts and novices, I show that experts (novice) generate mobile reviews that contain a slight (large) increase in emotional language and are more (less) attribute dense. Drawing on these findings, I advance managerial strategies for review platforms and service providers, and provide avenues for future research.

Co-Authorship Statement

I hereby certify that I am the principal contributor to and author of this dissertation. Chapter 2, at time of dissertation submission, is under second round review at the *Journal of Consumer Research*, and is co-authored with Xin (Shane) Wang, Xi Li, and June Cotte. Chapter 3 has not yet been submitted to a journal for publication. For both chapters, I was responsible for leading all aspects of the project including developing the theoretical bases of the paper; designing, collecting, and analyzing the experimental data; analyzing the field data; and writing the manuscript.

Acknowledgements

I would like to thank a few special people who have provided me with inspiration, guidance, and support over the course of my doctoral study. First and foremost, I would like to express my deep and sincere gratitude to my advisor, Shane. Thank you, Shane, for your mentorship; you taught me that research is not only a science, but also an art, where content and methodology is important, but so much more is required to produce high quality research. What I have learned from you over the past years is invaluable. Your vision, dynamism, and sincerity have deeply inspired me. It was a privilege and honour to have studied under your guidance. I would also like to thank you for your friendship, empathy, and humour. Although my doctoral study has come to an end, I believe our friendship and collaboration will continue well into the future.

I would also like to express my most sincere gratitude to my advisor, June. Thank you for your mentorship and your contribution to my development as a scholar. Your precision in language has shaped my ability to effectively communicate in my presentations and writings. I sincerely appreciate having you as a role model. Your leadership at Ivey and in the academic field is truly inspirational. You taught me how to combine professional excellence with fun and compassion.

Finally, a special thank you to my parents, Mary and Paul. I am extremely grateful for all the sacrifices that you have made to give my siblings and me the life that we had growing up in Canada. Thank you for your continued support, love, and prayers. I dedicate my doctoral achievement to you.

Table of Contents

Abstract i

Summary for Lay Audience iii

Co-Authorship Statement v

Acknowledgements vi

Table of Contents vii

List of Tables ix

List of Figures x

List of Appendices xi

Chapter 1: Introduction 1

Overview of Essay 1 3

Overview of Essay 2 5

References 7

Chapter 2: Essay 1 10

Overview of the Literature 15

Theory and Hypotheses 18

Overview of Studies 21

Study 1: Qunar (Field Data) 22

Study 2A: Priming an Aspect of Reviewer Expertise: Rating Familiarity (Experiment) ... 30

Study 2B: Priming an Aspect of Reviewer Expertise: Attribute Number (Experiment) 33

Study 3: TripAdvisor (Field Data) 37

Study 4: Yelp (Field Data) 44

General Discussion 48

References 55

Chapter 3: Essay 2 62

Overview of the Literature 67

Hypotheses 70

Study 1: Qunar (Field Data) 74

Study 2: TripAdvisor (Field Data) 85

General Discussion 95

References 100

Chapter 4: Final Remarks 105

References 108

Appendix 109

Curriculum Vita 111

List of Tables

Table 1. Description of the Qunar, TripAdvisor, and Yelp Datasets 23

Table 2. Description of Variables 24

Table 3. Key Summary Statistics of Variables 25

Table 4. Description of the Qunar, TripAdvisor, and Yelp Datasets 76

Table 5. Description of Variables 77

Table 6. Key Summary Statistics of Variables 78

Table 7. Study 1 (Qunar): The Effects of Mobile on Review Favorability. 80

Table 8. Study 2 (TripAdvisor): The Effects of Mobile on Review Favorability. 89

Table 9. Study 2 (TripAdvisor): The Effect of Mobile and Expertise on Review Emotionality. 92

Table 10. Study 2 (TripAdvisor): The Effect of Mobile and Expertise on Review Attribute Density. 91

List of Figures

Figure 1. Polarity of Evaluations as a Function of Platform-Defined Reviewer Expertise... 28

Figure 2. Study 2A Results 32

Figure 3. Study 2B Results 36

Figure 4. Difference in Ratings Between Experts and Novices as a Function General Level of Service by Service Providers..... 42

List of Appendices

Appendix A. Stimuli for Study 2A ……109

Appendix B. Stimuli for Study 2B ……110

Introduction

In marketing, word-of-mouth (WOM) is the act of consumers providing information about products, services, brands, or companies to other consumers (Richins and Root-Shaffer 1988). The communication of such information on the internet (e.g., reviews, tweets, blog posts) is known as electronic word-of-mouth (eWOM) (Babić Rosario et al. 2016; Chevalier and Mayzlin 2006; Hennig-Thurau et al. 2004). EWOM has been an important topic in marketing because it reduces the information asymmetry between firms and consumers at a massive scale (Mishra, Hedide, and Cort 1998) and plays a major role in shaping consumer choice (Hu, Liu, and Zhang 2008).

With more than 4.3 billion consumers connected on the Internet as of 2019 (Kemp 2019), consumers are able to not only read others' consumption-related experiences, but also share their own at an unprecedented scale. For example, in the US, 82% percent of consumers reported that they sometimes or almost always read online customer ratings or reviews when buying something for the first time; this figure jumps to a staggering 96% when only looking at the millennial cohort (Smith and Anderson 2016). Yelp, a major business review platform, has observed an exponential increase in the number of reviews generated on their platform over the past decade, reaching a total of 184 million reviews as of March 2019 (Yelp 2019).

Consumers' incorporation of online reviews into their decision process has not gone unnoticed by firms, many of whom actively try to harness eWOM as a marketing tool (Floyd et al. 2014). Many businesses incorporate the eliciting, collecting, and displaying of online reviews as part of their marketing efforts to stimulate product sales. For example, Amazon, a major online retailer, has encouraged consumers to post product reviews since 1995 and as of 2019,

contains over 110 million reviews (Feedback Express 2019). Amazon's online product reviews are very popular and are considered to be one of the site's more effective features. Many service providers, such as hotels and restaurants, offer incentives to designated experts across various review platforms to get them to write high quality reviews about the service provider, in an attempt to increase traffic to the business (Stone 2014). Given that consumers are increasingly sharing and consuming reviews, and that businesses are offering incentives to many reviewers, particularly elite reviewers, understanding the nature of platform-designated expert reviewers has become an important topic in consumer research.

More recently, the device on which reviews are generated has become of particular interest to marketing researchers (e.g., Melumad, Inman, and Pham 2019; Ransbotham, Lurie, and Liu 2019). As of 2014, the amount of time consumers spent on mobile devices surpassed their time spent on desktop computers (Business Insider Intelligence 2016). In 2019, approximately 96% of the US population owned a smartphone (Pew Research Center 2019). Similarly, review platforms have seen an upward trend in mobile device usage. For example, Yelp observed an increase on their mobile application from 8 million unique monthly active users back in 2012 to 33 million unique monthly active users by 2019 (Yelp 2019). Given the ubiquity of mobile devices in the hands of consumers, and the increasing prevalence of mobile-generated reviews, understanding the effects of generating reviews on mobile (vs. desktop) devices has become important to service providers and review platforms.

The purpose of this dissertation is to study the role of reviewer expertise and mobile devices on online review platforms. I define *reviewer expertise* as the extent to which a reviewer (i) contributes to an online platform and (ii) generates high quality reviews (e.g., degree of elaboration and category knowledge, favourability judgments by readers). By *mobile reviews*, I

refer to review content generated on portable and interface-constraining devices, such as a smartphone or a tablet. Major underlying goals in studying these topics are to provide theoretical contributions to the areas of eWOM, expertise, and mobile marketing, as well as to advance managerial strategies undertaken by review platforms and service providers. In the rest of this introduction, I provide an overview of the two essays in this dissertation. For each essay overview, I provide (i) key research questions to be investigated, (ii) a preview of my answers to these questions, (iii) how the research contributes to theory, and (iv) how the research advances managerial strategies.

Overview of Essay 1

In Essay 1, I study expert reviewers on online review platforms. Specifically, I investigate the following research questions: (i) Do platform-designated 'expert' reviewers actually exhibit features of expertise, as defined in the scientific literature (e.g., Alba and Hutchinson 1987)? (ii) How does expertise in generating reviews affect rating evaluations? (iii) What drives the effect? (iv) What downstream consequences do expert ratings have for service providers, such as restaurants and hotels?

My main tested hypothesis is that greater expertise in generating reviews leads to greater restraint from extremes in evaluations. I argue that repetition of generating reviews facilitates processing (Einhorn and Hogarth 1981; Hoyer 1984) and elaboration (Mandler and Johnson 1981), and enhances the number of attributes implicitly considered in evaluations (Johnson and Mervis 1987), which reduces the likelihood of assigning extreme summary ratings. This restraint-of-expertise hypothesis is tested across three different review platforms (TripAdvisor, Qunar, and Yelp), shown for both ratings and review sentiment, and demonstrated both between

(experts vs. novices) and within reviewers (expert vs. pre-expert), ruling out a purely self-selection explanation. Two experiments replicate the main effect and provide support for the attributes-based explanation. The field studies demonstrate two major consequences of the restraint-of-expertise effect. (i) Expert (vs. novice) reviewers have less impact on the aggregate valence metric, which is known to affect page-rank (Ghose, Ipeirotis, and Li 2012) and consumer consideration (Vermeulen and Daphne 2009). (ii) 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.

This research contributes to the literatures of eWOM and expertise in terms of (i) demonstrating why and how consumers designated as experts on review platforms actually resemble the conceptual definition of expertise in the consumer behavior literature, (ii) providing theory to explain, and demonstrating empirically, why online expert reviewers create less polarizing reviews, differing from novices, and (iii) showing that reviewer expertise is an antecedent of the aggregate valence metric and demonstrating that expert (vs. novice) reviewers play a lesser role on shifting valence metrics over time.

The research in Essay 1 provides two important managerial implications. First, my research challenges the common business practice of active solicitation of expert reviewers (Stone 2014). I delineate when and how expert reviewers benefit and harm service providers. Second, my research brings to light the issue of adopting ratings scales with the same granularity for experts and novices, and then combining expert and novice ratings to form an aggregate valence metric. I suggest that review platforms should adopt different rating scales for their expert and novice users.

Overview of Essay 2

Building on my research focus on expert reviewers from the first essay, in Essay 2, I study the role of mobile devices and reviewer expertise on online review platforms. Specifically, I investigate the following research questions: (i) How and why does generating reviews on mobile (vs. desktop) devices affect the actual review content and favorability judgments by readers? (ii) How and why might mobile (vs. desktop) reviews vary by the review platform? (iii) How and why might generating reviews on mobile (vs. desktop) devices vary for expert and novice reviewers?

Because of the relatively constraining interface of mobile devices, reviewers focus on the overall gist of their experiences (Melumad et al. 2019) and write shorter mobile (vs. desktop) reviews (Burtch and Hong 2014; Ransbotham et al. 2019). And because review length can enhance the diagnostic value for readers (Mudambi and Schuff 2010), I argue and show that whether mobile (vs. desktop) reviews are deemed more or less favorable by readers largely depends on the level of reduction in review length from desktop to mobile reviews. I show that this explanation of review length reduction accounts for the different findings on mobile reviews from past research, which analyzes online reviews from different platforms (Burtch and Hong 2014; Ransbotham et al. 2019). I postulate, and provide some empirical evidence, that a likely proximal cause for why review platforms vary in their length reduction from desktop to mobile reviews relates to differences in the mobile software interfaces.

Further, I argue, based on Schema Theory (Axelrod 1973; Mandler 2014), that expert and novice reviewers adopt different "strategies" in generating shorter mobile reviews. Because of their review-writing experience, experts develop a review-writing schema, and compared to novices, place greater emphasis on consistency in various review aspects, including emotionality

of language and attribute coverage in their mobile reviews. For example, although mobile reviews have been found to contain more emotional language than desktop reviews (Melumad et al. 2019; Ransbotham et al. 2019), I demonstrate that this observation is mitigated for experts relative to novices. Although mobile (vs. desktop) reviews are shorter for both experts and novices (Burtch and Hong 2014), I find that experts (novices do not) "compensate" by generating mobile reviews that are more (less) attribute dense.

This research contributes to the literatures of eWOM, mobile marketing, and expertise, in terms of (i) disentangling nuances in the relationship between mobile reviews and consumer judgments of review favorability, particularly across and within review platforms, (ii) demonstrating and explaining the heterogeneity of mobile reviews as a function of reviewer expertise, and (iii) contributing to the research area on the diagnostic value of eWOM (Mudambi and Schuff 2010) by elucidating the relationship between review length and review attribute density on readers' favorability judgments of reviews.

In terms of managerial implications, my research in Essay 2 brings to light a degree of caution to both service providers and review platforms in the elicitation of reviews from consumers. Given that increasingly more reviews are being generated on mobile devices (Yelp 2019), it is important for businesses to be aware of potential (negative) consequences of mobile reviews – e.g., reviewers vary in their enhanced use of emotional language and vary in their attribute coverage, which can affect the perceived diagnostic value to review-reading consumers. However, my research provides strategies to help address concerns about mobile-generated reviews.

References

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

Axelrod, Robert (1973), "Schema Theory: An Information Processing Model of Perception and Cognition." *American Political Science Review*, 67 (4), 1248-66.

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.

Burtch, Gordon, and Yili Hong (2014), "What Happens When Word of Mouth goes Mobile?" *Thirty Fifth International Conference on Information Systems, Auckland*.

Business Insider Intelligence (2016), "Mobile Apps are still Dominating Users' Time," http://www.businessinsider.com/mobile-apps-are-still-dominating-users-time-2016-9

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.

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.

Feedback Express (2019), "Amazon has 1,029,528 New Sellers this Year (Plus Other Stats)", https://www.feedbackexpress.com/amazon-1029528-new-sellers-year-plus-stats/

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.

Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh, and Dwayne D. Gremler (2004), "Electronic Word-of-Mouth via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?" *Journal of Interactive Marketing*, 18 (1), 38-52.

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

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.

Kemp, Simon (2019), "Digital 2019: Global Internet Use Accelerates," *We Are Social*, https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates

Mandler, Jean Matter. *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Psychology Press, 2014.

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

Melumad, Shiri, J. Jeffrey Inman, and Michel Tuan Pham (2019), "Selectively Emotional: How Smartphone Use Changes User-Generated Content," *Journal of Marketing Research*, 56 (2), 259-75.

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.

Pew Research Center (2019), "Mobile Fact Sheet", https://www.pewinternet.org/fact-sheet/mobile/

Smith, Aaron and Monica Anderson (2016), "Online Reviews", *Pew Research Center,* https://www.pewinternet.org/2016/12/19/online-reviews/

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

Ransbotham, Sam, Nicholas H. Lurie, and Hongju Liu (2019), "Creation and Consumption of Mobile Word of Mouth: How Are Mobile Reviews Different?" *Marketing Science*, 1-20.

Richins, Marsha L., and Teri Root-Shaffer (1988), "The Role of Evolvement and Opinion Leadership in Consumer Word-of-Mouth: An Implicit Model Made Explicit." *ACR North American Advances*.

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 2019, "An Introduction to Yelp Metrics as of March 31, 2019," https://www.yelp.ca/factsheet

Essay 1

Expert Reviewers' 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., sommelier, movie critiques). Over the past couple of decades, we have seen the rise of online reviews, where consumers not only rely on others' consumption-related experiences, but also share their own. Online review platforms now recognize their top users as 'expert' reviewers. For example, Yelp has its 'Elite' status, TripAdvisor has its 'Contributor Level', Qunar has its 'Expertise 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 so-called 'expert' reviewers has become an important topic in consumer research.

The study of online expert reviewers is particularly important for *service providers*, such as hotels and restaurants. Many businesses incentivize, by quite literally wining and dining, online expert reviewers, in order to get them to write high quality reviews for the business (Stone 2014). The underlying assumption is that having reviews written by expert reviewers ultimately helps the business. Therefore, a very important managerial question is whether this assumption is (always) true. If not, why and how might online expert reviewers not actually benefit, but actually harm, businesses? This is an important concern for many of today's service providers.

Understanding online expert reviewers is also critical for *review platforms*, such as TripAdvisor and Yelp. A major goal of online review platforms is to (accurately) capture 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 expert and novice reviewers can shape how various aspects of the platform are designed in order to more accurately capture and display past customer experiences.

Although substantial research has been conducted on online reviews (Babić Rosario et al. 2016; Floyd et al. 2014; You, Vadakkepatt, and Joshi 2015), we surprisingly know very little about online expert reviewers. Past researchers have examined experts across various domains, including 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; Parr, Heatherbell, and White 2002; Solomon 1990). Features of expertise include reduced cognitive effort (automaticity), enhanced cognitive structure (domain-specific knowledge), greater degree of elaboration, and enhanced memory for domain-related content (Alba and Hutchinson 1987; Ericsson and Smith 1991). Research highlights the importance of domain familiarity and practice in the development of expertise (Alba and Hutchinson 1987). Research in marketing has studied the nature of consumer expertise (Alba and Hutchinson 1987; Bettman and Sujan 1987) and the source credibility of expert recommendations on consumer choice (Biswas, Biswas, and Das 2006; Chocarro and Cortiñas 2013; Harmon and Coney 1982; Karmarker and Tormala 2009).

Although extant research has investigated various areas of expertise, little research has been conducted on *expert reviewers*. The domain of expert reviewers is novel because of the dual writer-reader characteristic of its users, its extremely large-scale nature, and its lack of formal qualifying tests to designate expertise levels. Given the prominent role of online reviews in shaping consumer choice, and the impact online reviews have on many of today's businesses,

expert reviewers on review platforms are an important marketing topic. Many questions about expert reviewers remain unanswered. I address the following issues and questions in my research: First, it is unclear whether online 'expert' reviewers actually exhibit features of *expertise*, as defined in the scientific literature (e.g., Alba and Hutchinson 1987). Second, how does expertise in generating reviews affect rating evaluations? Third, what drives the effect? Finally, what downstream consequences do expert ratings have for service providers, such as restaurants and hotels?

To answer these research questions I conduct three field studies (TripAdvisor, Qunar, and Yelp) and two experiments. Across the three review platforms, I find that so-called 'expert' reviewers actually do display features of expertise, including greater degree of elaboration, greater category knowledge, and greater perceived review favorability by readers. And although some platforms, such as Qunar and TripAdvisor, operationalize their 'expert' reviewers predominantly in terms of volume of past reviews generated, I find that the quantity-based approach still captures expertise. I acknowledge the lack of perfection in a predominantly quantity-based approach in capturing expertise; however, given the abundance of users and reviews, a predominantly quantity-based approach enables the relatively quick and scalable designation of expertise levels.

To be clear, the focus of my research is on the relationship between reviewer expertise and review ratings/content, so although in my analyses I do include some measures of consumer perceptions (e.g., 'Like', 'Helpful' and 'Useful' votes), it is not my intention to fully elucidate the perceptions of review-reading consumers on expert-generated reviews, but to focus on the effects of expertise on consumer-generated reviews.

In my research, I define *reviewer expertise* as the extent to which a reviewer (i) contributes to an online review platform – measured by number of past generated reviews written by the reviewer – and (ii) generates high quality reviews – measured across a number of dimensions, including degree of elaboration, degree of category knowledge, and review favorability judged by readers. My main hypothesis is that greater expertise in generating reviews leads to greater restraint from extremes in evaluations. The rationale is that repetition of generating reviews facilitates processing (Einhorn and Hogarth 1981; Hoyer 1984) and elaboration (Mandler and Johnson 1981), and enhances the number of attributes implicitly considered in evaluations (Johnson and Mervis 1987). 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), the consideration of larger numbers of attributes in evaluations reduces the likelihood of assigning extreme summary ratings.

This restraint-of-expertise hypothesis is tested and observed across three different review platforms, shown for both assigned ratings and review sentiment, and demonstrated not only between reviewers (experts vs. novices), but also within reviewers (expert vs. pre-expert), ruling out a purely self-selection explanation. 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) Expert (vs. novice) reviewers have less impact on shifting the aggregate valence metric, which is important, because valence metrics are known to affect service-provider page rank (Ghose, Ipeirotis, and Li 2012) and consumer consideration (Luca 2016; Vermeulen and Daphne 2009). (ii) Experts systematically benefit and harm service providers with their ratings. For service providers that generally provide mediocre

(excellent) experiences, expert reviewers assign significantly higher (lower) ratings than their novice counterparts.

My research provides two important managerial implications. First, the research challenges the common business practice of active solicitation of expert reviewers (Stone 2014). I delineate when and how expert reviewers benefit and harm service providers. Second, the research brings to light the issue of adopting ratings scales with the same granularity for experts and novices, and then combining expert and novice ratings to form an aggregate valence metric. I suggest that review platforms should adopt different rating scales for their expert and novice users. An in-depth discussion on the managerial implication of this research is provided later in the discussion section.

This essay makes three key contributions. First, the research bridges the gap between the topic of online expert reviewers and the more general literature on expertise (e.g., Alba and Hutchinson 1987). I provide empirical evidence that online expert reviewers do indeed exhibit features of traditional expertise, including a greater degree of elaboration and greater category knowledge.

Second, very little is known about the relationship between expertise and rating patterns. I provide conceptual and empirical support for the idea that greater expertise in generating reviews leads to less polarizing ratings, which is driven by the number of attributes considered by reviewers in their evaluations.

Third, although much of the extant research on online reviews provides support for 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. 2017). My research uncovers one such antecedent. I show that based on their rating approach, expert (vs.

novice) reviewers play a lesser role in shifting the aggregate valence metric. This finding complements and refines the conventional notion that expert recommendations highly affect consumer choice (Biswas, Biswas, and Das 2006; Chocarro and Cortiñas 2013). 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 aggregate valence metric over time means that experts (vs. novices) have a less important role in shaping the service providers that consumers will consider before reading individual reviews.

The rest of the paper is organized as follows. I first present a review of the background literature on online reviews and reviewer expertise, followed by my proposed hypotheses, which are based on existing psychological theory. Next, I present my five studies (three field studies and two randomized controlled experiments). Lastly, I discuss my main findings and provide managerial implications for service providers and rating platforms.

Overview of the Literature

Online peer reviews have been a hot 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 their level of analysis, into two groups: aggregate- (e.g., Chevalier and Mayzlin 2006; Moe and Trusov 2011; Sonnier, McAlister, and Rutz 2011) and individual-review levels (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 the 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. 2017; Fisher, Newman, and Dhar 2018). Although much research has been conducted on the predictive nature of aggregate metrics, very little is known about their antecedents. For instance, are there specific types of reviewers that tend to shift the existing aggregate valence metrics more (that is, who assign 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 review research, the unit of analysis is the individual review. The research in this area examines 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 a bit surprising that we actually know very little about reviewer expertise.

A few studies have touched on reviewer expertise on online peer review platforms (e.g., Liu and Park 2015; Racherla and Friske 2012). First, in most of these studies, reviewer expertise has only been operationalized in terms of number of past reviews generated by the reviewer. This operationalization is based on the assumption that greater experience in and familiarity with writing reviews enhances review-writing expertise. However, empirical evidence in support of this assumption is limited. Further, the studies do not clearly define reviewer expertise or test whether so-called 'expert' reviewers are actually experts, as defined in the scientific literature. For instance, the literature on expertise highlights various dimensions of expertise, including greater elaboration and greater domain-specific knowledge (Alba and Hutchinson 1987). Do platform-designated expert reviewers actually display some of these features? This question has not yet been investigated. Past studies have examined the relationship between reviewer expertise and review favorability by readers. The findings have been somewhat mixed, with some studies finding a positive correlation (Racherla and Friske 2012; Vermeulen and Seegers 2009; Zhang, Zhang, and Yang 2016) and other studies finding no correlation (Cheung, Lee, and Rabjohn 2008; Liu and Park 2015). Nonetheless, these research studies on expert reviewers are important, as they provide preliminary results for the study of reviewer expertise. As we begin to better understand the nature of expert reviewers, various gaps and questions remain to be addressed, including: How does expertise in generating reviews affect rating evaluations? If an effect exists, what drives it? What downstream consequences does the effect have for businesses? Given the limited research on expert reviewers and the increasing engagement businesses are having with expert reviewers (Stone 2014), it is critical for firms to understand the nature of expert reviewers in the online user-generated content domain.

Theory and Hypotheses

Repetition and Expertise

A major concern regarding so-called 'expert' 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 'expert' reviewers is first required. To define their 'expert' reviewers, review platforms generally assess their reviewers' quality (e.g., inclusion of photo/video, review elaboration, review favorability by readers) and quantity of reviews (number of past reviews generated). For most review platforms, such as Qunar and TripAdivsor, 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 a milestone of points moves a reviewer up along the expertise level designation. 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 expertise designation and a 'Community Manager' decides on whether or not that reviewer receives the official expertise badge; Yelp Support Center 2019).

Across most, if not all, review platforms, the common criterion of 'expertise' is generating lots of reviews. In other words, platform-defined 'expert' reviewers have a lot of experience and familiarity in generating reviews. Extant research on expertise highlights the importance of practice and familiarity 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, I predict that platform-defined 'expert' reviewers actually do display expertise features, such as greater review elaboration, greater domain-specific knowledge, and greater review favorability by readers.

H0: Reviewers who generate more reviews display greater degrees of expertise in their reviews.

Expertise and Rating Patterns

An important research question about expert reviewers is how expertise in generating reviews affects rating evaluations, if at all. Given that repetition of generating reviews is a common criterion in operationalizing reviewer expertise, and that repetition facilitates processing (Einhorn and Hogarth 1981; Hoyer 1984) and elaboration (Mandler and Johnson 1981), I predict that with greater experience in generating reviews, reviewers come to implicitly consider more domain-specific attributes (e.g., price, environment, location, cleanliness, and service) in their evaluations (Johnson and Mervis 1987). 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), I predict that the consideration of larger numbers of attributes in evaluations reduces the likelihood of assigning extreme summary ratings. I acknowledge that the assignment of extreme ratings can and do occur across all reviewers. However, I argue that the assignment of extreme ratings generally requires 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. I hypothesize 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.

Much research on online reviews has highlighted the importance of the aggregate valence metrics. A major finding is that aggregate valence metrics are predictive of firm sales (Babić Rosario et al. 2016; You, Vadakkepath, and Joshi 2015). Aggregate valence 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 (Luca 2016; Vermeulen and Daphne 2009). Although extant research has demonstrated the consequences of aggregate metrics (Floyd et al. 2014), very little is known about its antecedents. Because rating averages, by their nature, are generally skewed away from extreme values (Dai et al. 2017), I expect that as a natural consequence of their less polarizing rating approach, expert (vs. novice) reviewers have less impact on shifting aggregate valence metrics over time.

H3: Expert (vs. novice) 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, respectively)? I suspect novices' impact on the aggregate valence metrics is directional, and dependent on the general level of service by the service provider. The idea here is that based on the restraint-of-expertise hypothesis, relative to expert reviewers, novice reviewers adopt a more polarizing approach (i.e., an "I love it" vs. "I hate it" mentality). When presented with a positive experience, novice users 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 users, who are hesitant to give an extreme positive rating. 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 users, who are hesitant to give an extreme negative rating. Therefore, I hypothesize:

H4: For service providers that generally provide mediocre (excellent) experiences, experts assign significantly higher (lower) ratings than novices.

Overview of Studies

In this section, I present five research studies (three field studies and two experiments) investigating my hypotheses. Whereas the advantage of the three field studies is the generalizability – to the real world and across platforms – of the restraint-of-expertise hypothesis and its downstream consequences, the added value of the two experiments is in the provision of evidence for the causal inference and attributes-based explanation for my phenomenon of interest.

Study 1: Qunar (Field Data)

Purpose. The main goals of Study 1 are to test whether so-called 'expert' reviewers, as defined by the platform, actually display features of expertise, and to examine the relationship between reviewer expertise and rating polarity.

Variables and Analyses. In Study 1, I collect 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).

The main independent variable of interest is *reviewer expertise*, which is the extent to which a reviewer (i) contributes to an online review platform – measured by number of past generated reviews – and (ii) generates high quality reviews – measured across a number of dimensions, including degree of elaboration and review favorability by readers. In this study, I operationalize reviewer expertise based on Qunar's platform-defined *1-7 Expertise Level*. As previously mentioned, Qunar measures its expert reviewers using a point-based system on quality (e.g., inclusion of photos/videos) and quantity of reviews (number of past reviews generated). I used the natural logarithm of Qunar's *1-7 Expertise Level*, i.e., ln(*Expertise_level*), in my analysis to normalize its distribution. Throughout the analyses, I provide descriptive statistics for the first two Expertise Levels, levels 1 and 2, and the last two Expertise Levels, levels 6 and 7.

In order to test whether platform-defined expertise is consistent with the general literature definition of expertise (e.g., Alba and Hutchinson 1987), I test the relationship of Qunar's *1-7 Expertise Level* with a number of expertise-related dimensions, including review quantity (the number of past reviews generated by the reviewer), review elaboration (the number of Chinese characters used in the review) and review favorability (the number of 'Like' votes received by the review).

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

	Qunar (Study 1)	**TripAdvisor (Study 3)**	**Yelp (Study 4)**
Language	Chinese	English	English
Number of Cities	4	6	4
List of Cities	Beijing, Gaungzhou, Sanya, Shanghai	Chicago, HK, London, Los Angeles, Paris, Singapore	Las Vegas, Phoenix, Pittsburgh, Toronto
Service Provider Type	Hotel	Hotel	Restaurant
Number of Service Providers per City	15	10	50
Total Number of Service Providers	60	60	200
Number of data points (i.e., individual reviews)	125,985	39,203	49,380
Date of Data Collection	March 2016	January 2017	January 2018
Dates of Reviews	Oct 2007 – Mar 2016	Feb 2016 – Jan 2017	May 2005 – Dec 2017

Notes:
Qunar & TripAdvisor:
Reviews from Qunar and TripAdvisor were scrapped from their online website: https://www.qunar.com/ and https://www.tripadvisor.ca/
Selection of hotels were based on popularity on the platform at the time of data scraping. While I collected and analyzed all the review data available in the selected hotels on Qunar, I only collected and analyzed the most recent 1 year of review data on TripAdvisor.
Yelp:
Yelp review data was compiled from the data provided by Kaggle.com: https://www.kaggle.com/yelp-dataset/yelp-dataset Two groups of data were compiled: by restaurant and by reviewers. The by-restaurant review data, shown in the above table, was collected to test H4.
Specific cities were selected based on having the most number of restaurants listed. Fifty restaurants from each city were randomly selected. The by-reviewer data was collected to test H0, H1, and H2. The by-reviewer data consisted of over 1 million reviews. The detailed reviewer information allowed me to categorize each review as having being generated by a pure novice (i.e., has never been elite), a pre-elite, or an elite reviewer.

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 is the review.
MonthsAgo	Number of months ago review was posted at the 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|.
Reviewer	Identification of reviewer; only included in Yelp analysis. Treated as random effect in the mixed models.
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.

Table 3. Key Summary Statistics of Variables

	Qunar (Study 1) N = 125,985				**TripAdvisor** (Study 3) N = 39,203				**Yelp** (Study 4) N = 49,380			
	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	1.3	3.0	0	207
Length	83.8	185.9	1	7,306	586.4	514.6	86	8,605	656.2	605.1	12	5,005
MonthsAgo	14.4	8.0	0	101	6.9	3.2	1	12	39.4	28.8	1	150
Rating	4.46	0.91	1	5	4.33	0.95	1	5	3.74	1.33	1	5
RatingPolarity	1.61	0.62	0	2	1.49	0.67	0	2	1.34	0.72	0	2
ReviewerExpertise	1.52	0.88	1	7	2.53	2.07	0	6	0.27	0.45	0	1

Impact of rating on the aggregate valence metrics 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, I 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, I conduct my 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 'Expert' Reviewer (H0). To test whether Qunar's platform-defined 'expert' reviewer designation is consistent with the literature-defined concept of expertise (Alba and Hutchinson 1987), I examine how various expertise-related features of reviews vary as a function of Qunar's platform-defined expertise levels. Consistent with H0, I find that reviewers higher on Qunar's *1-7 Expertise Level* (i) have generated more reviews ($M_{Levels\_1\_2}$ = 3.3 vs. $M_{Levels\_6\_7}$ = 35.1 past reviews, r = .84, p < .001), (ii) 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 (iii) 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).

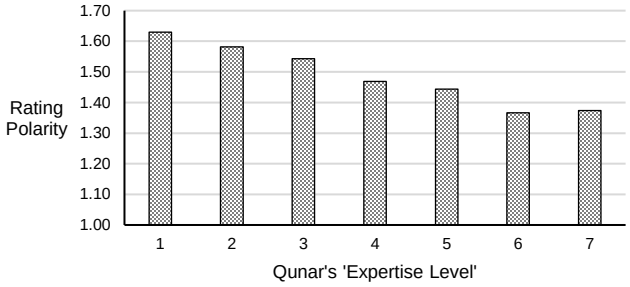
(ii) Expertise and Rating Evaluations (H1). Next, I test the relationship between reviewer expertise and rating polarity. In accordance with H1, results from my linear mixed-effects regression model show that reviewers higher on Qunar's *1-7 Expertise Levels* demonstrate greater restraint from extremes 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; see **Figure 1A**). As a robustness test, I relax my parametric assumption about the rating polarity dependent variable by conducting an ordered logistic regression (using *polr()* function in the *MASS* package in R; Ripley et al. 2013); my restraint-of-expertise results are robust (β = -0.33, t = -24.55, p < .001).

I conduct another robustness analysis comparing the dispersion of ratings 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).

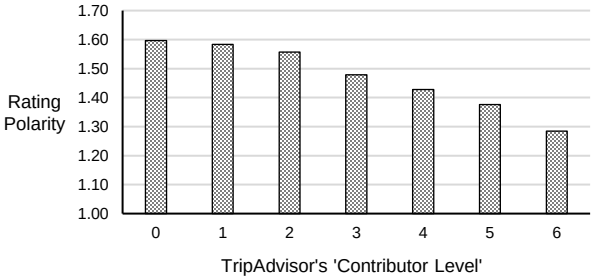
My explanation for the restraint-of-expertise effect is based on attributes implicitly considered by reviewers when making their overall rating evaluation (H2). Later, in my English-based review field data, I algorithmically detect and count the number of category-related nouns

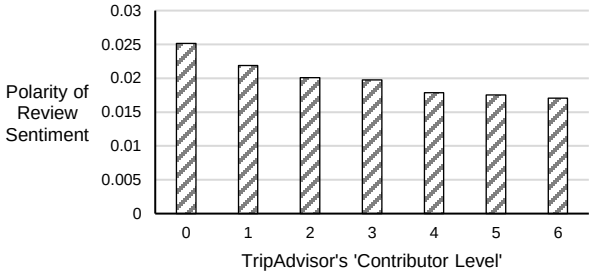
Figure 1. Polarity of Evaluations as a Function of Platform-Defined Reviewer Expertise.

A) **Qunar** (Study 1)



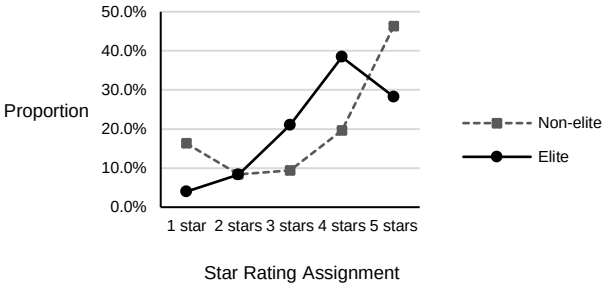
B) **TripAdvisor** (Study 3)

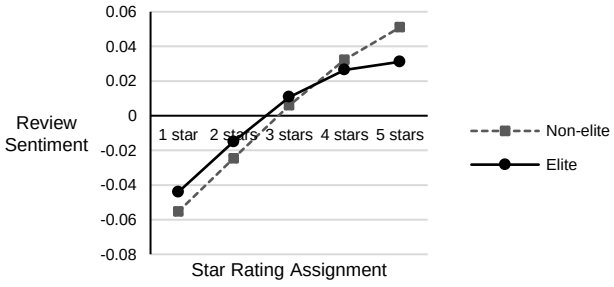




Review sentiment calculated using the LIU sentiment-word dictionary (Liu 2012).

C) **Yelp** (Study 4)





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

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, I am unable to extract the specific attributes mentioned in the reviews. I 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), I test the mediating role of review length on the restraint-of-expertise effect. 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 total restraint-of-expertise effect. That is, experts consider more attributes, which leads to a less extreme, or restrained, overall evaluations.

(iii) Impact of expertise on shifting aggregate valence metric (H3). Next, I test the impact of expertise on the aggregate valence metric. Consistent with H3, the results from my mixed-effects model demonstrate a significant negative effect of reviewer expertise on the impact on aggregate valence metric – both in terms of the moving valence metric ($M_{Level\_1\_2}$ = 0.63 vs. $M_{Level\_6\_7}$ = 0.56; β = -0.48, $t(124870)$ = -8.90, $p < .001$) and the cumulative valence metric ($M_{Level\_1\_2}$ = 0.67 vs. $M_{Level\_6\_7}$ = 0.57; β = -0.50, $t(125916)$ = -5.28, $p < .001$).

Conclusions. In Study 1, using Qunar hotel review data, I demonstrate that platform-defined 'expert' reviewers certainly do exhibit features of expertise, including greater review elaboration and greater reader-assessed review favorability (H0). This finding highlights the value of a predominantly quantity-based approach, as used on Qunar, in capturing reviewer expertise. I show that expert (vs. novice) reviewers adopt a less polarizing rating approach (H1), which appears to be in part driven by how much they consider in their evaluations (H2). As a consequence, experts have less impact on shifting aggregate valence metrics (H3), which is

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 my case, reviewer expertise and less polarizing rating evaluations. Is the observed phenomenon driven purely by a self-selection bias? For example, reviewers that do not write reviews often (i.e., novice reviewers) might only do so when experiences are either extremely good or extremely bad. Or is the relationship also causal in nature, such that as reviewers generate more reviews, their reviews, both in terms of assigned ratings and review sentiment, become more restrained?

I speculate that, to some degree, both a self-selection bias and a causal relationship are present in the restraint-of-expertise effect. In subsequent studies – Studies 2A, 2B, and 4 – I test and demonstrate the causal effect of expertise on less polarizing rating evaluations. I conduct randomized controlled experiments in Studies 2A and 2B, where I manipulate aspects of reviewer expertise – rating familiarity and considered attributes – to test the effect of reviewer expertise on less polarizing rating evaluations. Later in Study 4, by analyzing Yelp restaurant reviews, I further test and provide evidence for the effect of reviewer expertise on less polarizing rating evaluations by tracking, intra-reviewer, how the polarity of assigned ratings and review sentiment change as reviewers generate more reviews.

Study 2A: Priming an Aspect of Reviewer Expertise: Rating Familiarity (Experiment)

Purpose. The purpose of Study 2A is to test the effect of reviewer expertise on the polarity of rating evaluations. Given that a key criterion, across more-or-less all review platforms, in

operationalizing their expert reviewers is the number of past reviews generated, in Study 2A, I test whether priming a key aspect of reviewer expertise – rating familiarity – affects the polarity of rating evaluations.

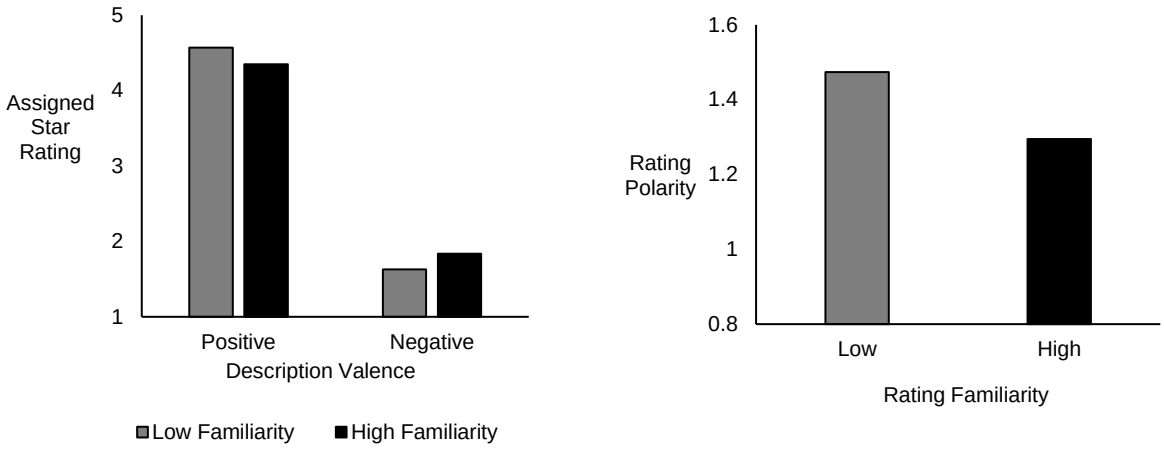
Design. The design of the experiment is a 2 rating familiarity (high vs. low) x 2 description valence (positive vs. negative) between-subjects design. The outcome measure in the experiment is the assigned star rating, along a 5-point scale from 1-*Terrible* to 5-*Excellent* (see **Appendix A** for experimental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

Procedure. Online participants (N = 190, $\%_{female}$ = 56.3%, M_{Age} = 35.0, SD_{Age} = 11.1) on Amazon Mechanical Turk were randomly assigned to the high or low rating familiarity condition. Participants assigned to the high [low] rating familiarity condition were asked to think about and rate three restaurants they have visited [electronic products they have purchased] over the past year (note that the dependent measure is specific to restaurants). 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 experience.

Results. A two-way analysis of variance revealed a significant main effect of description valence ($M_{positive}$ = 4.46 vs. $M_{negative}$ = 1.74; F(1,186) = 912.93, p < .001) and no main effect of rating familiarity (expertise) on assigned star rating (*ns*). As expected, the interaction between rating familiarity (expertise) and description valence on assigned star rating is significant ($F(1,186)$ = 5.68; p = .018; see **Figure 2**).

A follow-up analysis shows that for the positive experience description, participants primed with high rating familiarity assigned marginally lower ratings (M = 4.35) than those primed with low rating familiarity (M = 4.57; t(1,94) = 1.89, p = .06). For the moderately negative experience

Figure 2. Study 2A Results



description, there was no significant difference in ratings between the high and low rating familiarity groups ($M_{high\_familiarity}$ = 1.84 vs. $M_{low\_familiarity}$ = 1.63, *ns*).

Next, I looked at the polarity rating variable, my main dependent variable. Consistent with my prediction, I find that participants primed with high rating familiarity (a dimension of expertise) assigned ratings that were less polarizing (M = 1.29 average units from the midpoint of a five-point scale) than those primed with low rating familiarity (M = 1.47; $t(185) = 2.12$, p = .035).

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

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

Purpose. The purpose of Study 2B is to further test the effect of reviewer expertise on the polarity of rating evaluations. Given my theorizing that expert reviewers consider more attributes in their evaluations, which drives the restraint-of-expertise effect, I 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, rating along attributes are done *before* assigning an overall rating. I test how the number of attributes considered might affect the overall rating. Consistent with H2, I 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 (2 vs. 6) x 2 experience valence (positive vs. negative). The outcome measure in the experiment is the assigned star rating, along a 5-point scale from 1-*Terrible* to 5-*Excellent* (see **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 (2 vs. 6) conditions. Participants were first asked to rate the recent restaurant experience across either two or six 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 rating of the experience. All ratings were assigned along a 5-point 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.

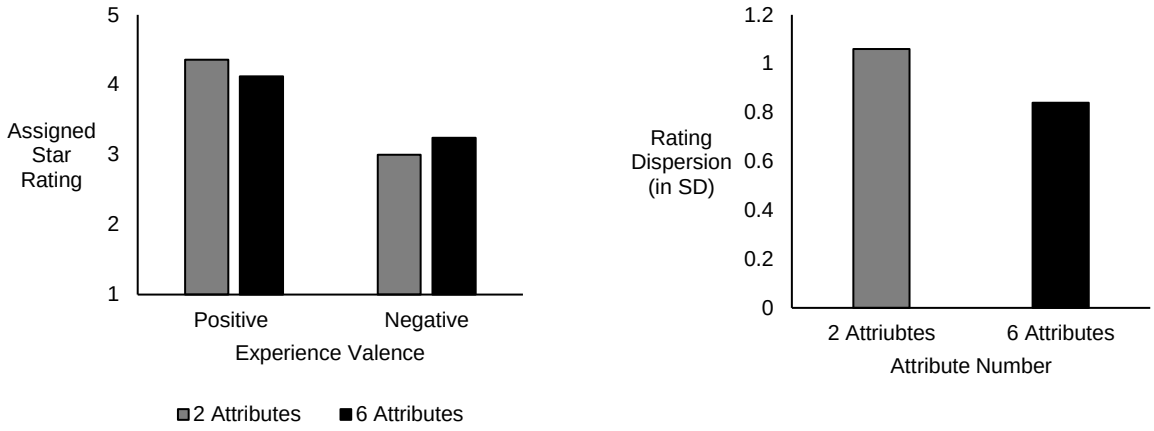
Results. As an attention check, I removed participants that were asked to report a positive (negative) restaurant experience, but reported an experience rating of 1-star (5-stars). This lead to the removal of 24 of the 240 data points, bringing the total participant count to 216.

A two-way ANCOVA revealed a significant main effect of experience valence ($M_{positive}$ = 4.23 vs. $M_{negative}$ = 3.13, $F(1,207) = 113.58$, $p < .001$) and no main effect of number of attributes on assigned star rating (*ns*). As expected, the interaction between experience valence and attribute number on the assigned star rating was significant ($F(1,207) = 4.49$, $p = .035$; see **Figure 3**). (Controls in the ANCOVA included 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). Following up on the interaction, I 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; $t(111) = 2.19$, $p = .03$). 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, I 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 3**).

As a robustness analysis, I also test the polarity of ratings based on the distance of the ratings from the average rating across all participants. I 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; $t(214) = 2.27$, $p = .024$).

Figure 3. Study 2B Results



Conclusion. Across Studies 2A and 2B, I demonstrate two different ways, related to expert reviewers – rating familiarity and the number of considered attributes – that can reduce the rating polarity. These findings provide support for the causal relationship between reviewer expertise and restraint ratings. Further, results from the Qunar field data (Study 1), demonstrate the generalizability of the phenomenon in the real-world.

Although considerable support for the restraint-of-expertise phenomenon has been provided, 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? (iii) Does the attenuated impact of ratings by 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 study.

Study 3: TripAdvisor (Field Data)

Purpose. In Study 3, I test whether the restraint-of-expertise effect, H1, as observed in reviews from the Chinese-based platform Qunar.com, (i) replicates in a North American-based platform, TripAdvisor.com, and (ii) is also exhibited in the sentiment of written reviews. Further, I 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.

Variables and Analyses. In Study 3, I collected and analyzed over 39,000 online reviews, over a one year time span, of hotels from TripAdvisor.com, a major online English-based travel review platform (see **Table 1** for description of dataset; see **Table 2** for variable list; see **Table 3** for summary statistics of variables).

The main independent variable of interest is *reviewer expertise*. I operationalize reviewer expertise based on TripAdvisor's platform-defined *0-6 Contributor Level*. Similar to Qunar, TripAdvisor measures their expert reviewers using a points-based system on quality (e.g., inclusion of photos/videos) and quantity of reviews (number of past reviews generated). I used the natural logarithm of TripAdvisor's *0-6 Contributor Level*, i.e., ln(*Contributor_level* + 1), in my analysis to normalize its distribution. Throughout the analyses, I provide descriptive statistics for the first two Contributor Levels, levels 0 and 1, and the last two Contributor Levels, levels 5 and 6.

In order to test whether platform-defined expertise is consistent with the general literature definition of expertise (e.g., Alba and Hutchinson 1987), I test the relationship of TripAdvisor's *0-6 Contributor Level* with a number of expertise-related dimensions, including review quantity (the number of past reviews generated by the reviewer), review elaboration (the number of characters and words used in the review), category knowledge (the number of category-related attributes in the review), and review favorability (the number of 'Helpful' votes received by the review).

A key moderating variable I test is *general level of service* by the business, which is operationalized in this study by a moving user rating average, based on most recent 20 reviews prior to generating the review.

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 on these variables, see Study 1). I also compare the relative assigned ratings between experts and novices. Because the reviews on TripAdvisor are in English, I was 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). (I 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 the dataset, I only kept the nouns. Next, I removed city-specific terms by conducting term frequency-inverse document frequency (*tf-idf*) analysis across the six cities. This allowed me to compile 30 of the most frequently used hotel-related nouns; e.g., *service, location*, and *view.* Next, for each review, using a match and count based algorithm, I identified the number of unique nouns mentioned in the review that were contained in the list of 30 hotel-related nouns. This produced my number of hotel-specific attributes mentioned in each review.

Because there are multiple reviews of each hotel, that is, the reviews are nested within hotels, I conduct my main analyses with mixed effects regressions, with maximum likelihood

estimation. Included in the analyses are a number of control variables, including hotel ID (as a random effect), date of review post (converted to number of months from date of review scraping), expertise level of the prior reviewer posting about the service provider (to control for some interdependencies amongst reviewers), and purpose of travel (transformed to five dummy variables).

Results: *(i) Platform-Defined 'Expert' Reviewer (H0)*. Consistent with H0, I find that reviewers higher on TripAdvisor's *0-6 Contributor Level* exhibit features of expertise, in terms of (i) having generated more reviews ($M_{Levels\_0\_1}$ = 1.6 vs. $M_{Levels\_5\_6}$ = 114.4 past reviews, r = .93, p < .001), (ii) having a higher degree of elaboration in their reviews (by number of characters: $M_{Levels\_0\_1}$ = 430.8 vs. $M_{Levels\_5\_6}$ = 740.2, r = .34, p < .001; by number of words: $M_{Levels\_0\_1}$ = 71.7 vs. $M_{Levels\_5\_6}$ = 125.0, r = .26, p < .001), (iii) including a greater number of category-related attributes in their reviews ($M_{Levels\_0\_1}$ = 3.4 vs. $M_{Levels\_5\_6}$ = 5.0 hotel-related attributes considered in review, r = .25, p < .001), and (iv) 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). Next, I test whether expertise in generating reviews affects rating evaluations. Results from my 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, Ω^2 = 0.019; see **Figure 1B**). As a robustness test, I relax my parametric assumption about the rating polarity dependent variable by conducting an 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 another robustness analysis, I 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 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 = 308.65, p < .001).

Further, I test the restraint-of-expertise effect not only on the assigned ratings, but also on the sentiment of the review text. My results show that the restraint-of-expertise effect is also displayed in the polarity of the sentiment of the review text (by Bing-Liu's word-sentiment dictionary: $M_{Level\_0\_1}$ = 0.024 vs. $M_{Level\_5\_6}$ = 0.017, β = -0.004, t = -23.39, p < .001, see **Figure 1B**; by AFINN word-sentiment dictionary: $M_{Level\_0\_1}$ = 0.048 vs. $M_{Level\_5\_6}$ = 0.032, β = -0.008, t = -23.60, p < .001).

(iii) Mechanism: Attributes Considered (H2). I test whether the number of considered attributes drives the restraint-of-expertise effect. As a measure for the number of considered attributes, I use the number of domain-specific (hotel-related) nouns mentioned in the reviews, which was extracted using Part-of-Speech tagging. Using mediation analysis in R (*mediation* R package, Tingley et al. 2014), I find that number of considered attributes mediates the effect of reviewer expertise on less polarizing ratings (-0.0035, 95% CI: -0.0044 to -0.0026, 1000 iterations).

(iv) Impact of expertise on shifting the aggregate valence metric (H3). Next, I test the impact of expertise on aggregate valence metrics. My results demonstrate that 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; β = -0.06, $t(39115)$ = -13.96, p <

.001) and the cumulative valence metric ($M_{Level\_0\_1}$ = 0.73 vs. $M_{Level\_5\_6}$ = 0.62; β = -0.07, $t(39136)$ = -17,74, p < .001).

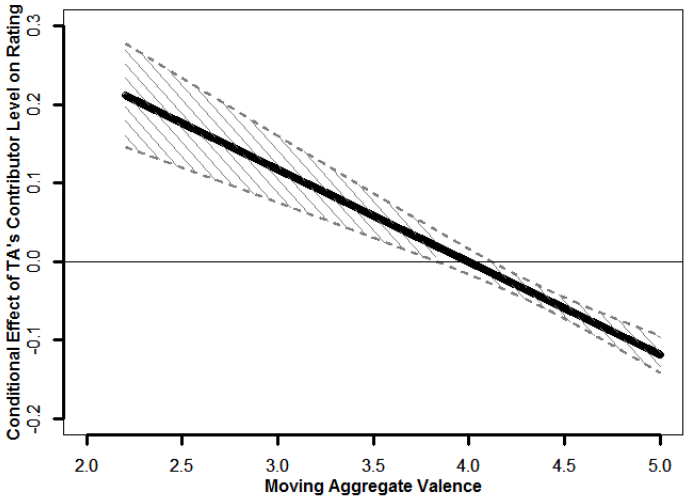
(v) Relative ratings between experts and novices (H4). Lastly, I test the relative ratings between expert and novice reviewers and how they might depend on the general level of service provided by the business. Results from my mixed-effects regression model show that there is a significant interaction between the general level of service and TripAdvisor's measure of reviewer expertise on assigned ratings (β = -0.11, t(39113) = -7.34, p < .001; see **Figure 4A**).

Given that I am interested in detecting focal values of general level of service where experts (vs. novices) assign systematically higher and lower ratings, I conduct a follow-up floodlight analysis (Johnson and Neymar 1936; Spiller et al. 2013). My floodlight analysis demonstrates that for service providers that generally provide mediocre to poor experiences (specifically, recent average ratings below 3.8, see **Figure 4A**), 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). For service providers that generally provide excellent experiences (specifically recent average ratings above 4.1), 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).

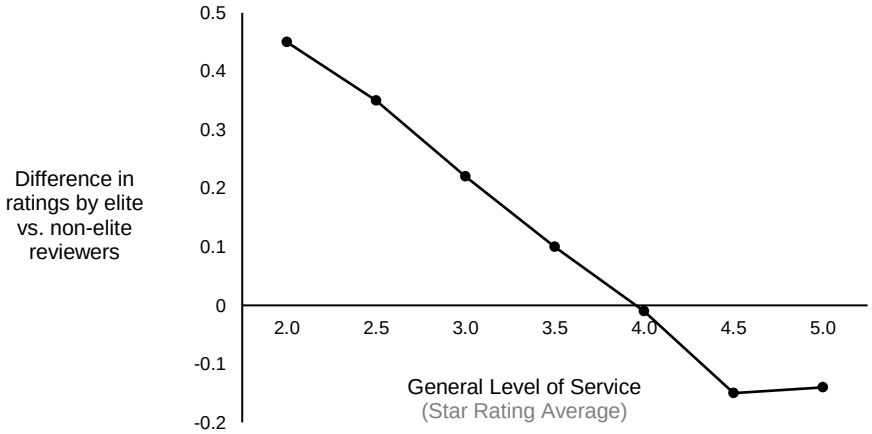
Conclusions. In this study, using hotel reviews from TripAdvisor, I replicate the restraint-of-expertise effect, evidenced not only in the assigned ratings, but also the sentiment of the review text. Further, I demonstrate two major consequences of the restraint-of-expertise effect. First, expert (vs. novice) reviewers have less impact on the aggregate valence metric. Second, I demonstrate that expert (vs. novice) reviewers 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.

Figure 4. 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)



Although I have provided evidence to support the restraint-of-expertise hypothesis, the two field studies presented so far have only demonstrated the effect *between* reviewers (i.e., experts vs. novices). This is in part because reviewers were collected for the same service providers, rather than the same reviewers. In Study 4, I address this shortcoming by collecting and analyzing reviews *by reviewers* (expert vs. pre-expert), as well as by service providers. I also test whether the findings from the previous two field studies are replicated.

Study 4: Yelp (Field Data)

Purpose. The main purpose of Study 4 is to test the restraint-of-expertise effect, H1, not only between reviewers (experts vs. novices), as was tested and evidenced in the previous two field studies, but also within reviewers (experts vs. pre-experts). I also test whether the systematic beneficial and harmful impact of expert (vs. novice) ratings, H4, as demonstrated in Study 3's TripAdvisor hotel reviews, replicates in Study 4's Yelp restaurant reviews.

Variables and Analyses. For Study 4, I collected and analyzed online Yelp restaurant reviews, a major online restaurant review platform based in North America (see **Table 1** for description of dataset; see **Table 2** for variable list; see **Table 3** for summary statistics of variables).

The main independent variable is *reviewer expertise*. I operationalize reviewer expertise based on Yelp's platform-defined 'Elite' status designation. As stated on 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' makes a decision 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 reviewer 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 me to conduct my *within reviewer* analyses, where I compare and contrast reviews generated before and after a reviewer obtained her first 'Elite' badge.

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

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

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

Results: *(i) Platform-Defined 'Expert' Reviewer (H0)*. From my *between* reviewer (expert vs. novice) analyses, I find that Yelp 'Elite' (vs. Yelp non-'Elite') reviewers demonstrate greater features of reviewer expertise, in terms of (i) having generated more reviews (M_{Elite} = 226.9 vs. $M_{Non\text{-}elite}$ = 13.3 past reviews, r = .56, p < .001), (ii) having a higher degree of elaboration in their reviews (by characters per review: M_{Elite} = 919.9 vs. $M_{Non\text{-}elite}$ = 554.4, r = .32, p < .001; by words

per review: M_{Elite} = 174.1 vs. $M_{Non\text{-}elite}$ = 105.6, r = .32, p < .001), (iii) including a greater number of domain-specific (restaurant) attributes in their reviews (M_{Elite} = 8.1 vs. $M_{Non\text{-}elite}$ = 5.5 restaurant attributes mentioned in reviews, r = .24, p < .001), and (iv) having generated reviews that are deemed more favorable by readers (M_{Elite} = 2.4 vs. $M_{Non\text{-}elite}$ = 0.8 average 'Useful' votes per review post, r = .32, p < .001).

My *within* reviewer (expert vs. pre-expert) analyses involves examining only reviews from users who have obtained the Yelp 'Elite' badge. I compare and contrast reviews that were generated prior to, versus after, 'Elite' badge designation. In line with my between reviewer results, I 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} = 919.9 vs. $M_{Pre\text{-}elite}$ = 664.0, r = .14, p < .001; by words per review: M_{Elite} = 174.1 vs. $M_{Pre\text{-}elite}$ = 126.2, r = .14, p < .001), greater number of domain-specific attributes mentioned in the reviews (M_{Elite} = 8.1 vs. $M_{Pre\text{-}elite}$ = 6.4 restaurant attributes per review, r = .10, p < .001), and greater degree of favorability by readers (M_{Elite} = 2.4 vs. $M_{Pre\text{-}elite}$ = 1.1 average 'Useful' votes per review post, r = .13, p < .001).

(ii) Expertise and Rating Evaluations (H1). In line with results from the previous field studies and experiments, I find evidence for the restraint-of-expertise hypothesis between expert and novice reviewers when comparing by rating polarity (M_{Elite} = 1.11 vs. $M_{Non\text{-}elite}$ = 1.53 average distance from midpoint of 5-point scale; β = -0.57, t = -279.2, p < .001, Ω^2 = .07; see **Figure 1C**) as well as by variance in ratings (SD_{Elite} = 1.08 vs. $SD_{Non\text{-}elite}$ = 1.52; K^2 = 35,630, p < .001). More importantly, I observe the restraint-of-expertise effect *within* expert reviewers (by rating polarity: M_{Elite} = 1.11 vs. $M_{Pre\text{-}elite}$ = 1.22; β = -0.16, t = -35.09, p < .001; and by variance in ratings: SD_{Elite} = 1.08 vs. $SD_{Pre\text{-}elite}$ = 1.21; K^2 = 424.1, p < .001).

As a robustness analysis of H1, I also test whether expert, versus novice, reviewers express more restraint in the sentiment of their review text. Indeed, results show that the review sentiment by expert (vs. novice) reviewers is less polarizing, even when controlling for the assigned rating (β = -0.02, t = -28.63, $p < .001$, Ω^2 = .02; see **Figure 1C**).

(iii) Mechanism: Attributes Considered (H2). Regarding H2, I test whether the number of considered attributes drives the restraint-of-expertise effect. As a measure of the number of considered attributes, I 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), I find that number of considered attributes mediates the restraint-of-expertise effect, in both my *between* reviewers (-0.0351, 95% CI: -0.0399 to -0.0303, 1000 iterations, 13.6% proportion of main effect mediated) and *within* reviewers analyses (-0.0411, 95% CI: -0.0518 to -0.0304, 1000 iterations, 16.3% proportion of main effect mediated).

(iv) Relative ratings between experts and novices (H4). I test who – expert or novice reviewers – give better ratings and how it might depend on the general level of service provided by the business. Results from my 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, $p < .001$; see **Figure 4B**).

Specifically, I see that for restaurants with 2.0, 2.5, 3.0, and 3.5 average star ratings, experts, on average, assigned *higher* ratings than novices by 0.45, 0.35, 0.22, and 0.1, respectively (all p's < .001). In contrast, for restaurants with 4.5 and 5.0 average star ratings, experts assign *lower* ratings than novices by 0.15 and 0.14, respectively (both p's < .001).

Conclusion. Using restaurant reviews from Yelp, I demonstrate the restraint-of-expertise effect (H1), shown for both assigned ratings and review sentiment, and demonstrated both between reviewers (experts vs. novices) and within reviewers (experts vs. pre-experts). I provide evidence for the mechanism of number of attributes considered (H2). Finally, I 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, I study experts on online review platforms. My main hypothesis is that greater expertise in generating reviews leads to greater restraint in rating evaluations. Across five studies (three field studies and two experiments), I 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 tested and observed across three different review platforms (TripAdvisor, Qunar, and Yelp), shown for both ratings and review sentiment, and demonstrated not only between reviewers (experts vs. novices), but also within reviewers (expert vs. pre-expert), ruling out a purely self-selection explanation. Two experiments replicate the main effect and provide support for the attributes-based explanation. The field studies demonstrate two major consequences of the restraint-of-expertise effect. (i) Expert (vs. novice) reviewers play a lesser role in shifting the aggregate valence metric over time. (ii) 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 important theoretical implications of my work. First, my research extends the reach of the literature on expertise to the online user-generated content (UGC) domain. Much of extant research on expertise has been 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; Parr, Heatherbell, and White 2002; Solomon 1990). However, given the rise of UGC and the ability of UGC platforms to differentiate amongst its 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. For scalability, many platforms simply implement a point-based system to designate their expert users, where users receive points for the quantity and quality of their contributions to the platform and certain milestones of points designate a particular expertise level. So, are these so-called online 'expert' users really experts, as defined in the scientific literature (e.g., Alba and Hutchinson 1987)? My research suggests that the answer is 'generally yes'. I acknowledge the lack in the perfection in capturing expertise with a points-based approach, especially one that places heavier weight on quantity over quality; however, I concede that such an approach is practically reasonable given the large-scale nature of many UGC platforms. Future research can work on refining efficient and scalable approaches that more effectively capture expertise.

Second, my 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., computer, high-definition television) versus low technology-oriented products (e.g., treadmill, mattress). Packard and Berger (2017) show that novices are more likely to use explicit endorsement styles in the reviews (e.g., "I recommend it" vs. "I like it"), which are found to be more persuasive and increase purchase intent. The researchers suggest that *ceteris paribus*, the endorsement styles novices and experts tend to use can lead to greater persuasion by novices. In my research, I demonstrate how the restraint-of-expertise effect can dampen the influential nature of experts. Because experts generally assign ratings that are less polarizing, *in the context where information is abundant and aggregated*, experts have less impact on shifting the aggregate valence metric, which affects 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 experts may be more favored by consumers (Racherla and Friske 2012; Zhang, Zhang, and Yang 2016), the attenuated impact experts have on aggregate valence metric over time means that experts (vs. novices) play a mitigated role on the service providers consumers consider before reading individual reviews.

My research has three important practical implications for business. First, my research challenges the notion of companies actively seeking and incentivizing expert reviewers. I delineate when and how expert reviewers benefit and harm service providers. Service providers that generally provide *excellent* levels of service should avoid expert reviewers, 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. In my data I find that whereas experts most frequently assign 4-star ratings, novices most frequently assign 5-star ratings (see **Figure 1C**). As a consequence, I find that service providers that generally provide

excellent levels of service receive lower ratings from experts than from novices, and therefore, benefit more from novices in terms of elevating their user rating average.

Service providers that generally provide *mediocre* service can benefit from reviews by experts. Relative to experts, novices adopt a more polarizing rating approach. I find that novices assign more 1-star ratings (17%) than 2-star ratings (9%), but the opposite is true for expert reviewers (3% 1-star ratings versus 9% star-ratings), who rarely assign 1-star ratings, even after controlling for the service provider. As a consequence, I find that service providers that generally provide *mediocre* levels of service receive lower ratings from novices than from experts, and therefore, benefit more from experts in terms of elevating (or not further lowering) their user rating average.

Second, an important concern for many online platforms is the type of rating scale – binary (thumbs up/down) or multiple point (5-star or 10-point) – they should adopt. A key criterion in selecting the appropriate rating scale is to select one where its users can and do evaluate along a similar level of granularity. A scale that is relatively too coarse may miss out on detailed differences, and a scale that is relatively too fine is inefficient and may lack rating consistency. Consider the example of YouTube. In the early years, the company used a 5-star rating scale. YouTube came to realize that the 5-star rating scale was inefficient, as almost all ratings were either 1 or 5 stars. As a result, in 2010, the company decided to switch to using a thumbs up/down rating scale (Rajaraman 2009). I suspect that the type of scale that should be adopted by a platform depends on (i) the relative comparability of the content being evaluated (e.g., similar hotel experiences vs. diverse types of videos) and (ii) the expertise of the evaluators on the platform.

(i) I speculate that if the content being evaluated is relatively comparable (e.g., experiences at restaurants or hotels), then with rating practice, users are more likely to be able to discern nuances across the similar content, and adopt a more granular rating approach. However, if the evaluated content is relatively diverse (e.g., videos or music varying in length, content, and style), users are less likely to develop an implicit reference frame to evaluate the diverse content, and therefore adopt a more polarizing rating approach.

(ii) Results from my research show that whereas expert evaluators are more likely to adopt a restrained rating approach, novice evaluators are more likely to adopt a polarizing rating approach. Therefore, a recommendation for platforms is to implement two different rating scales for their expert and novice evaluators. Interestingly, this is actually what is already done on Rotten Tomato, where their 'critic' (expert) reviewers evaluate along a 10-point scale and their 'audience' (novice) evaluators rate along a 5-point scale.

Last, my research brings to light the issue of combining expert and novice ratings to form a single aggregate valence metric. The combining of their ratings to form a single aggregate valence metric would not be problematic *if their rating averages were more or less similar*. However, as we can see in **Figure 4**, this not the case – expert and novices assign systematically different ratings depending on the general level of service of the business. As a result, I recommend platforms implement two separate aggregate valence metrics, one for ratings by their experts and the other for ratings by their novices. This additional information can be highly valuable and informative to consumers who may prefer rating averages of experts over novices, or vice versa. Interestingly, this approach too has already been adopted by Rotten Tomato where aggregate metrics of ratings are separated for their 'audience' (novice) and 'critic' (expert) reviewers.

This research paves the way for a number of future research projects on reviewer expertise. First, as discussed, I believe future research can study and establish more efficient scalable approaches that more effectively capture expertise. For example, are there other important criteria other than quantity of reviews generated that should be used by review platforms in their operationalization of expertise? In designating reviewer expertise, how does the transparent point-based system, as used by TripAdvisor and Qunar, compare to alternative systems, such as the nomination system adopted by Yelp? A reasonable place to start to answer these questions is by studying reviewer expertise across different platforms, comparing the different criteria and measurement systems, and assessing their effectiveness in capturing expertise, as defined in the literature.

Second, much of this research focused review content/ratings as a function of reviewer expertise; little attention was paid to motivations of expert and novice reviewers. Extant research highlights various reasons for why consumers generate and share their product/service experiences (Berger 2014; Hennig-Thurau et al. 2004; Packard and Wooten 2013). Hennig-Thurau et al. (2004) propose that consumers engage in online word-of-mouth because of their desire for social interaction, their desire for economic incentives, their concern for other consumers, and the potential to enhance their own self-worth. Given these various reasons, how and why might expert and novices reviewers vary in their motivation to share product/service experiences? To what degree? How might the motivations to engage in eWOM for expert and novice reviewers affect their review content and ratings? These are some important questions for future research.

Finally, the focus of my research is on the relationship between reviewer expertise and review content/rating. Although my 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 experts? What role does the expertise badge (e.g., ‘Elite 2019’) 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? I believe these are some important questions that remain to be answered in the area of reviewer expertise.

To conclude, this research provides evidence, in the context of user-generated review platforms, of how expertise in generating reviews 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.

References

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

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.

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

Bettman, James R., and Mita Sujan (1987), "Effects of Framing on Evaluation of Comparable and Noncomparable Alternatives by Expert and Novice Consumers," *Journal of Consumer Research*, 14 (2), 141-54.

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.

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 (2017), "Aggregation of Consumer Ratings: An Application to Yelp.com," *Quantitative Marketing and Economics*, 16 (3), 289-339.

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.

Ericsson, K. Anders, and Jacqui Smith, eds. (1991), *Toward a General Theory of Expertise: Prospects and Limits*, Cambridge University Press.

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

Floyd, Kristopher, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling (2014), "How Online 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 Nonmobile Reviews on Consumer Purchase Intentions," *Journal of Marketing Research*, 1-18.

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.

Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh, and Dwayne D. Gremler (2004), "Electronic Word-of-Mouth via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?" *Journal of Interactive Marketing*, 18(1), 38-52.

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

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

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.

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 Psychologyi* 9 (1), 111-51.

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.

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 David B. Wooten (2013), "Compensatory Knowledge Signaling in Consumer Word-of-Mouth," *Journal of Consumer Psychology*, 23 (4), 434-50.

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

Parr, Wendy V., David Heatherbell, and K. Geoffrey White (2002), "Demystifying Wine Expertise: Olfactory Threshold, Perceptual Skill and Semantic Memory in Expert and Novice Wine Judges," *Chemical Senses*, 27 (8), 747-55.

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*.

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."

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 2019, 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.

Essay 2

The Differential Effects of Generating Reviews on Mobile Devices for Expert and Novice Reviewers

A major current trend is consumers' increasing use of mobile devices. As of 2014, the amount of time consumers spent on mobile devices surpassed their time spent on desktop computers (Business Insider Intelligence 2016). In 2016, approximately 88% of the US population owned a smartphone, with a staggering 98% smartphone ownership within the millennial cohort (Nielsen 2016). Similarly, review platforms have seen an upward trend in mobile device usage. For example, Yelp, a major business review platform, observed an increase on their mobile application from 8 million unique monthly active users back in 2012, to 33 million unique monthly active users by 2019; no meaningful change has been observed on their desktop website over the same timeframe (Yelp 2019). Given the ubiquity of mobile devices in the hands of consumers and the increasing prevalence of mobile-generated reviews, understanding the effects of generating reviews on mobile (vs. desktop) devices, as well as its heterogeneity across reviewers and review platforms, has become an important topic to marketing researchers (e.g., Melumad, Inman, and Pham 2019; Ransbotham, Lurie, and Liu 2019).

The topic of mobile-generated reviews is particularly important to *review platforms*, such as Yelp or TripAdvisor. Major goals of online review platforms are to increase the activity on their platforms and present (accurate) information on past customer experiences to prospective review-seeking customers. Given that mobile devices can and do facilitate the goal of increasing

activity on the review platform, it has become important for platforms to understand how generating reviews on mobile (vs. desktop) devices affect the actual review content, as well as readers' judgments of the reviews. This information is important as it can shape how various aspects of the platform's mobile application is designed for its users.

The study of mobile-generated reviews is also very important to *service providers*, such as hotels and restaurants. A major goal of service providers is to maintain an active online presence on a number of review platforms, such as Google, TripAdvisor, and Yelp, as research has shown the positive impact of review volume on firm performance (Babić Rosario et al. 2016; Duan, Gu, and Whinston 2008). To this end, many businesses encourage their patrons, in both offline and online ways, to write online reviews about their service experiences. In particular, many service providers offer incentives to designated experts across various review platforms to get them to write high quality reviews about the service provider, in an attempt to increase traffic to the business (Stone 2014). Given that an increasing number of reviews are generated on mobile devices (Yelp 2019) and that businesses are offering incentives to many reviewers, particularly elite reviewers, understanding the effect of generating reviews on mobile (vs. desktop) devices, and its heterogeneity across reviewer expertise, can help shape the review elicitation strategies adopted by service providers.

Although an abundance of research has been conducted on online reviews (Babić Rosario et al. 2016; Floyd et al. 2014; You, Vadakkepatt, and Joshi 2015), we still know very little about reviews generated on mobile devices. Recently, some research has been published on mobile-generated reviews (e.g., Melumad, Inman, and Pham 2019; Ransbotham, Lurie, and Liu 2019). A consistent finding across different review platforms is that reviews generated on mobile devices are a lot shorter in length than reviews generated on desktop devices (Burtch and Hong 2014;

Melumad et al. 2019; Ransbotham et al. 2019). Melumad et al. (2019) argue that reviews generated on mobile devices are shorter because of the physically constraining nature of mobile devices, which encourages reviewers to focus on the overall gist of their experience.

With regard to the review content, mobile (vs. desktop) reviews have been found to be more concrete (Ransbotham et al. 2019) and more emotional in language (Burtch and Hong 2014; Melumad et al. 2019). While Burtch and Hong (2014) argue that the enhanced emotional language in mobile reviews is afforded by the portability of smartphones – where reviewers are more likely to write mobile reviews in an irrational or emotional state (Loewenstein 1996) – Melumad et al. (2019) claim that the enhanced emotional language in mobile reviews is also influenced by the physically constraining nature of mobile devices – where focusing on the overall gist of an experience manifests in the greater use of emotional language.

From the perspective of the readers, there appears to be no clear consensus as to whether mobile reviews are deemed more or less favorable by readers than their desktop counterparts. For example, analyzing review data from Urbanspoon.com, Ransbotham et al. (2019) find that relative to desktop reviews, mobile reviews are *less* favored by readers. In contrast, analyzing review data from TripAdvisor.com, Burtch and Hong (2014) find the opposite – mobile (vs. desktop) reviews receive *more* favorability votes by readers.

Although these recently published papers provide a basis for understanding the relationship between mobile devices and generated review content, various gaps in our knowledge about mobile reviews remain to be addressed. First, it is unclear why extant research on mobile reviews find conflicting results on the favorability of mobile reviews by readers. Elucidating heterogeneity across platforms would provide deeper insights into the effects of generating review content on mobile (vs. desktop) devices. Second, an underlying assumption in past

research on mobile reviews is that the effects of generating reviews on mobile devices is relatively homogeneous across reviewers. However, it is unclear whether this actually is the case, given the differences shown between expert and novice reviewers in Essay 1. Third, although a key finding on mobile (vs. desktop) reviews is that they are significantly shorter due to the constraining interface of mobile devices (Melumad et al. 2019), little is known about whether different approaches are taken by reviewers when writing shorter mobile reviews. For example, in writer shorter mobile reviews, do reviewers compensate by writing reviews that are more attribute dense? If so, to what degree?

To address these gaps in our knowledge, I investigate the following research questions: (i) How and why does generating reviews on mobile (vs. desktop) devices affect the actual review content and favorability judgments by readers? (ii) How and why might mobile reviews vary by the review platform? (iii) How and why might generating reviews on mobile (vs. desktop) devices vary for expert and novice reviewers?

Because of the relatively constraining interface of mobile devices, reviewers focus on the overall gist of their experiences (Melumad et al. 2019) and write shorter mobile (vs. desktop) reviews (Burtch and Hong 2014; Ransbotham et al. 2019). And because review length can enhance the diagnostic value for readers (Mudambi and Schuff 2010) – that is, help the decision process by increasing consumer's confidence in their purchase decision – I argue and show that whether mobile (vs. desktop) reviews are deemed more or less favorable by readers largely depends on the level of reduction in review length from desktop to mobile reviews. I show that this explanation of review length reduction accounts for the different findings on mobile reviews from past research, which analyzes online reviews from different platforms (Burtch and Hong 2014; Ransbotham et al. 2019). I postulate, and provide some empirical evidence, that a likely

proximal cause for why review platforms vary in their length reduction from desktop to mobile reviews relates to differences in the mobile software interface.

Further, I argue, based on Schema Theory (Axelrod 1973; Mandler 2014), that expert and novice reviewers adopt different "strategies" in generating shorter mobile reviews. Because of their review-writing experience, experts develop a review-writing schema. I argue and show that compared to novices, experts place greater emphasis on consistency in various aspects of reviews, including emotionality of language and attribute coverage in their mobile reviews. For example, although mobile reviews have been found to contain more emotional language than desktop reviews (Melumad et al. 2019), I demonstrate that this observation is mitigated for experts relative to novices. Although mobile (vs. desktop) reviews are shorter for both experts and novices (Burtch and Hong 2014), I find that experts (novices do not) "compensate" by generating mobile reviews that are more (less) attribute dense.

The research in this essay provides three main contributions. First, this research disentangles nuances about the relationship between mobile reviews and consumer judgments of review favorability. This research highlights how the degree of reduction in review length from desktop to mobile is a major predictor about readers' favorability of mobile reviews and addresses cross-platform differences on mobile reviews. Second, this research demonstrates and explains the heterogeneity of mobile-generated reviews as a function of reviewer expertise. Specifically, I show that in their mobile (vs. desktop) reviews, relative to novices, experts include less enhanced emotional language and place greater emphasis on attribute coverage. Finally, this research contributes to the area of information diagnosticity of eWOM (Mudambi and Schuff 2010) by elucidating the relationship between review length and review attribute density on readers' favorability judgments of reviews. I show that although review attribute density has a positive

effect on readers' favorability judgments of reviews, this effect is particularly pronounced for shorter reviews.

The rest of the paper is organized as follows. I first present a review of the background literature on online reviews, followed by my proposed hypotheses, which are based on existing psychological theory. Next, I present my two field studies. Lastly, I discuss my main findings, practical implications, and limitations to my research.

Overview of the Literature

Online reviews have been an important topic in marketing over the past decade. They reduce the information asymmetry between firms and consumers (Mishra, Hedide, and Cort 1998) and have played a major role in shaping consumer choice (Hu, Liu, and Zhang 2008). Firms have become more attentive to the impact of online reviews (Floyd et al. 2014). Much of the research on online reviews have studied the impact of reviews on (i) firm sales (Chevalier & Mayzlin 2006; Floyd et al. 2014) and (ii) consumer opinion (Mudambi & Schuff 2010; Peng et al. 2014).

Researchers have studied the effects of online reviews' aggregate measures, such as volume (number of reviews) and valence (user rating averages). A major finding in this area is that aggregate metrics are predictive of firm sales (Babić Rosario et al. 2016). For example, Duan, Gu, and Whinston (2008) find that the volume of online review postings has a significant effect in predicting box office sales. Chevalier & Mayzlin (2006) find that improvements in valence of book reviews leads to an increase in relative sales on Amazon and Barnes & Noble. More recently, a few meta-analyses have been published examining the overarching relationships between aggregate metrics and firm performance (Babić Rosario et al. 2016; Floyd et al. 2014; You, Vadakkepath, and Joshi 2015).

The relationship between online reviews and consumer opinion has also been examined. Analyzing individual reviews, researchers have shown that review length has a positive effect on how favorable consumers find the reviews (Liu and Park 2015; Mudambi & Schuff 2010; Peng et al. 2014), which is driven by the information diagnosticity of longer reviews (Mudambi and Schuff 2010). Researchers have shown that review texts that are more readable (Korfiatis, Garcia-Bariocanal, and Sanchez-Alonso 2012; Liu and Park 2015), contain more negative sentiment (Ludwig et al. 2013; Peng et al. 2014), use anxious (vs. angry) tone (Yin et al. 2014), and include either highly subjective or highly objective content (but not a mix) (Ghose and Ipeirotis 2011), are more likely to influence consumer opinion. The specific descriptions in reviews can influence consumer attitudes (Moore 2015; Packard & Berger 2017). Moore (2015) shows that for utilitarian products, explained actions ("I *chose* this product because…") are favored by readers, whereas for hedonic products, explained reactions ("I *love* this product because…") are more favorable. Packard and Berger (2017) find that compared to reviews with implicit endorsements (e.g., "I liked it"), reviews with explicit endorsements (e.g., "I recommend it") are more persuasive and increase purchase intent.

Reviewer characteristics have also been found to be important in online reviews. For example, the disclosure of reviewer identity enhances how helpful readers find the review post, which is driven by message persuasiveness (Forman, Ghose & Wiesenfeld 2008; Ghose and Ipeirotis 2011; Kusumasondjaja, Shanka, and Marchegiani 2012). The reputation, or number of friends, of the reviewer has a positive effect on credibility of the review (Racherla and Friske 2012). In Essay 1, I examined how and why the expertise of consumers in generating reviews shapes their rating evaluations, and the downstream consequences this has on aggregate valence metrics. I argued and showed that greater expertise in generating reviews leads to greater

restraint from extremes in evaluations, which is driven by the number of attributes considered in the review.

More recently, the device on which reviews are generated have become of particular interest to marketing researchers (e.g., Melumad, Inman, and Pham 2019; Ransbotham, Lurie, Liu 2019). A consistent finding in this area is that mobile (vs. desktop) reviews are a lot shorter (Burtch and Hong 2014; Melumad et al. 2019; Ransbotham et al. 2019). Melumad et al. (2019) argue that the physically constraining interface of mobile devices (e.g., small keyboard and screen) encourages reviewers to focus on the overall gist of their experience, and hence, write shorter reviews.

Further, Melumad et al. (2019) find that focusing on the gist tends to manifest as reviews that emphasize the emotional aspects of an experience rather than more specific details. Burtch and Hong (2014) also find that mobile (vs. desktop) reviews contain more emotional language, but attribute this finding to the portability of mobile devices, where reviewers are more likely to generate reviews closer in time to the consumption experience and are more likely to be in an irrational, emotional state (Loewenstein 1996).

Mixed findings have been observed on readers' favorability of mobile (vs. desktop) reviews. For example, analyzing reviews from Urbanspoon, Ransbotham et al. (2019) find that mobile reviews are deemed *less* favorable by readers than desktop reviews. In contrast, analyzing TripAdivsor reviews, Burtch and Hong (2014) find that opposite – mobile (vs. desktop) reviews are judged to be *more* favorable.

Grewal and Stephen (2019) argue and show that the label on the review post indicating whether or not the review was generated on a mobile device (e.g., "via mobile") can also affect consumer opinion. Grewal and Stephen contend that mobile reviews are deemed more accurate

by readers due to the belief that writing reviews via mobile requires more effort and effort translates to the reviewer being more trustworthy.

Although these recently published papers provide a basis for understanding the relationship between mobile devices and generated review content, as highlighted in the introduction, various gaps about mobile reviews remain to be addressed. It is unclear (i) why extant research on mobile reviews find conflicting results on the favorability of mobile reviews by readers, (ii) whether the effects of generating reviews on mobile devices vary across reviewers, such as expert and novice reviewers, and (iii) whether different approaches are taken by reviewers when writing shorter mobile reviews.

Hypotheses

Favorability of Mobile Reviews across Review Platforms

Given that mobile reviews have consistently been found to be shorter than desktop reviews (Burtch and Hong 2014; Melumad et al. 2019) and that review length plays an important role in providing diagnostic value to readers (Mudambi and Schuff 2010), I hypothesize that the degree of reader favorability of mobile (vs. desktop) reviews largely depends on the level of reduction in review length from desktop to mobile reviews.

H1: The favorability of mobile (vs. desktop) reviews by readers depends on the level of reduction in review length from desktop to mobile reviews.

Researchers have arrived at opposing conclusions about the favorability of mobile reviews by readers (e.g., Burtch and Hong 2014; Ransbotham et al. 2019). What might account for the mixed findings? Assuming H1 is true, one might expect that different conclusions on the favorability of mobile reviews have been drawn because researchers have analyzed mobile (vs.

desktop) reviews from different platforms and there may be considerable variation in the reduction of review length from desktop to mobile reviews across different platforms. Therefore, analyzing mobile (vs. desktop) reviews from a platform with a relatively large (e.g., 60%) reduction in review length from desktop to mobile reviews, one is more likely to observe an overall *negative* effect of mobile on review favorability.

Past research on mobile reviews show that there are aspects of mobile reviews that will tend to *increase* reader favorability judgments of the reviews. For example, Melumad et al. (2019) show that mobile reviews contain more emotional language and greater use of emotional content increases persuasion (Ludwig et al. 2013). Grewal and Stephen (2019) show that when readers know a review is generated from a mobile device, as indicated by the mobile-generated label (e.g., "via mobile") on the review post, readers perceive the review to have required a greater amount of effort to write, which enhances how trustworthy and accurate readers view the review, and therefore enhances readers' favorability judgments of the mobile-generated review.

Putting these findings together with the general reduction in review length of mobile (vs. desktop) reviews, we can conclude that there are two general "forces" of mobile reviews, where the reduction in review length when generating reviews on mobile (vs. desktop) devices has a negative effect on review favorability (Mudambi and Schuff 201), and other aspects of mobile reviews, such as the enhanced use of emotional language (Melumad et al. 2019) and readers' knowledge of the review being generated on a mobile device (Grewal and Stephen 2019), have a positive effect on review favorability. Therefore, analyzing mobile (vs. desktop) reviews from a platform with a relatively small (e.g., 10%) reduction – i.e., minimizing the negative effect of mobile review length on favorability – one is more likely to observe an overall *positive* effect of mobile on review favorability.

Mobile Reviews and Reviewer Expertise

Recent research shows that mobile (desktop) reviews contain more emotional language (i.e., words conveying affective content, independent of valence; Burtch and Hong 2014; Melumad, Inman and Pham 2019). Melumad et al. (2019) argue that generating content on mobile (vs. desktop) devices leads consumers to generate brief content, which encourages them to focus on the overall gist of their experience. They demonstrate that the focus on gist, in turn, leads to the selective reporting of affective information, yielding content that is more emotional. Burtch and Hong (2014) argue that mobile devices afford consumers increased opportunities to access the internet, enabling impulsive, emotional actions, which would otherwise subside if reviewers were required to wait before taking action (Ariel and Loewenstein 2006; Loewenstein 1996; Loewenstein 2000). In summary, there are two mechanisms – constraining interface and portability of mobile devices – that lead reviewers to use more emotional language when generating reviews on mobile (vs. desktop) devices. An underlying assumption is that this emotional effect of generating reviews on mobile (vs. desktop) devices is relatively homogeneous across reviewers. However, it is unclear whether this actually is the case, given the differences shown between expert and novice reviewers in Essay 1. In Essay 1, I demonstrated that compared to novices, experts have greater restraint from extremes in their rating evaluations and use less emotional language in their reviews. Therefore, this begs the question, is the enhanced use of emotional language in mobile (vs. desktop) reviews (Burtch and Hong 2014; Melumad et al. 2019) consistent between novice and expert reviewers?

Although generating reviews on mobile (vs. desktop) devices influences reviewers to use more emotional language, based on Schema Theory, I predict that relative to novices, experts are less affected by the general emotional influence of generating reviews on mobile (vs. desktop)

devices. Schema Theory proposes that all knowledge is organized into units, call schemata (singular: schema) (Axelrod 1973; Mandler 2014). Past experiences shape the development of schemata (Alba and Hutchingson 1987) and influences behaviour (e.g., driving a car, playing a sport) (Rentsch, Heffner, and Duffy 1994), in particular, the consistency of behaviour across a variety of context (Beilock and Carr 2001; Goldstein and Chance 1980; Ziefle 2002).

Because of their extensive review-writing experience, expert reviewers develop a review-writing schema, and therefore, compared to novices, are expected to be more consistent in various aspects of their reviews when generated on mobile and desktop devices. Therefore, I hypothesize that the enhanced use of emotional language on mobile devices is mitigated for experts relative to novices.

H2: The enhanced use of emotional language in mobile (vs. desktop) reviews is mitigated for experts relative to novices.

Past research consistently shows that reviewers write shorter reviews on mobile (vs. desktop) devices (e.g., Melumad et al. 2019; Ransbotham et al. 2019). However, little is known about whether different approaches are taken by reviewers when writing shorter mobile reviews. Schema Theory would suggest that experts aim to, at least implicitly, produce relatively consistent review content, regardless of contextual cues, including device type (Ziefle 2002). Therefore, given that both experts and novices write shorter mobile (vs. desktop) reviews, I predict that experts "compensate" when generating shorter mobile reviews by discussing a greater relative number of attributes in their mobile reviews. In other words, I predict that compared to novices, experts generate mobile (vs. desktop) reviews that are more attribute dense.

H3: Compared to novices, experts generate mobile (vs. desktop) reviews that are more attribute dense.

Review Attribute Density

Research on online reviews suggest that review length can provide diagnostic value to consumers (Mudambi and Schuff 2010), especially if the information can be obtained without additional search costs (Johnson and Payne 1985). Open-ended reviews provide additional explanations and context to the assigned star rating and can affect the perceived helpfulness of a review (Mudambi and Schuff 2010). In a similar way, review attribute density – that is, how many different attributes discussed in the review relative to its length – is also an important factor in effectively and efficiently conveying information to review-reading consumers. I predict that there is a positive effect of review attribute density on review favorability. However, consistent with theory of information overload (Jacoby 1974, 1984), I also predict that the positive effect of review attribute density is more pronounced for when reviews are shorter.

H4A: There is a positive effect of review attribute density on readers' review favorability judgments.

H4B: The positive effect of review attribute density (H4A) is more pronounced for shorter reviews.

Study 1: Qunar (Field Data)

The purpose of Study 1 is to investigate readers' favorability judgments of mobile-generated reviews. Specifically, I address the question of why mixed findings on the favorability of mobile reviews have been observed across different review platforms (e.g., Burtch and Hong 2014; Ransbotham et al. 2009). Further, I investigate how expert and novice reviewers may be differentially affected by generating reviews on mobile (vs. desktop) devices.

Dataset. In Study 1, I collect and analyze over 123,000 online reviews of hotels on Qunar.com, a major online travel review platform in China (see **Table 4** for description of dataset; see **Table 5** for variable list; see **Table 6** for summary statistics of variables). The dataset only includes reviews posted between January 2011 and December 2015; Qunar's mobile application was first introduced in 2011.

Variables. The main independent variable of interest is *mobile*, which is a binary variable indicating whether the review was generated on a mobile or desktop device. In the dataset, 92.1% of reviews are generated on mobile devices. Qunar also distinguishes amongst three types of mobile reviews – whether the review was generated on the mobile application, the short messaging service (SMS, also known as text messaging), or the mobile website – which make up of 59.3%, 30.4%, and 2.4%, respectively, of reviews in the dataset.

The moderating variable of interest is *reviewer expertise,* which is the extent to which a reviewer (i) contributes to an online review platform – measured by number of past generated reviews – and (ii) generates high quality reviews – measured across a number of dimensions, including degree of elaboration and review favorability by readers (from Essay 1). In this study, I operationalize reviewer expertise based on Qunar's platform-defined *1-7 Expertise Level*. Qunar measures its expert reviewers using a point-based system on quality (e.g., inclusion of photos/videos) and quantity of reviews (number of past reviews generated). I used the natural logarithm of Qunar's *1-7 Expertise Level*, i.e., ln(*ExpertiseLevel*), in my analysis to normalize its distribution. Descriptive statistics are provided for the first two *Expertise Levels*, levels 1 and 2, combined, which make up 85.4% of all reviews in the dataset, and the last three *Expertise Levels*, levels 5, 6, and 7, combined, which make up 1.0% of all reviews in the dataset. That is, there are many more novices than experts in the data.

Table 4. Description of the Qunar, TripAdvisor, and Yelp Datasets

	Qunar (Study 1)	**TripAdvisor (Study 2)**
Language	Chinese	English
Number of Cities	4	6
List of Cities	Beijing, Gaungzhou, Sanya, Shanghai	Chicago, HK, London, Los Angeles, Paris, Singapore
Service Provider Type	Hotel	Hotel
Number of Service Providers per City	15	10
Total Number of Service Providers	60	60
Number of data points (i.e., individual reviews)	123,529	99,050
Date of Data Collection	March 2016	January 2017
Dates of Reviews	Jan 2011 – Dec 2015	Jan 2012 – Dec 2016

Notes:
Qunar & TripAdvisor:
Reviews from Qunar and TripAdvisor were scrapped from their online website: https://www.qunar.com/ and https://www.tripadvisor.ca/
Selection of hotels were based on popularity on the platform at the time of data scraping.

Table 5. 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 is the review.
MonthsAgo	Number of months ago review was posted at the 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.*
ReviewerExpertise	Platform-defined reviewer expertise (Qunar = *1-7 Expertise Level*, TripAdvisor = *0-6 Contributor Level*)
ServiceProvider	Identification of hotel/restaurant to which the review is attributed. Treated as random effects in the mixed models.

Table 6. Key Summary Statistics of Variables

	Qunar (Study 1) N = 123,529				**TripAdvisor** (Study 2) N = 99,050			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Favorability	0.4	2.7	0	219	0.5	0.9	0	14
Length	83.8	185.9	1	7,306	586.4	514.6	86	8,605
MonthsAgo	14.4	8.0	0	101	6.9	3.2	1	12
Rating	4.46	0.91	1	5	4.33	0.95	1	5
ReviewerExpertise	1.52	0.88	1	7	2.53	2.07	0	6

My main dependent variable of interest is *review favorability*, which is operationalized by the number of 'Like' votes a review receives. For robustness of measurement, I also operationalize review favorability by the number of comments the review receives.

Analyses. Because my dependent variable of interest, review favorability, is a count variable, I conduct my main analyses using a Poisson regression model. Included in the analyses are a number of control variables, including hotel ID (*Hotel*), date of review post (converted to number of months from date of review scraping, *MonthsAgo*), and purpose of travel (transformed to five dummy variables, *Purpose*).

$$Favorability = \beta_0 + \beta_1\, Mobile + \beta_2 \ln(ExpertiseLevel) + \beta_3\, Mobile \times \ln(ExpertiseLevel) + \beta_5\, MonthsAgo + \beta_{6\text{-}10}\, Purpose + \beta_{11\text{-}70}\, Hotel + \varepsilon$$

Results: (i) Mobile → Favorability. Results from the Poisson regression model show that on the Qunar platform, there is an overall negative effect on the favorability of mobile (vs. desktop) reviews (M_{Mobile} = 0.28 versus $M_{Desktop}$ = 1.27 average 'Like' votes per review post; b = -0.695, se = 0.016, z = -44.223, p < .001; see **Model A1** in **Table 7**). Findings are robust for when including in the analyses only reviews that contain (i) one or more 'Like' votes (M_{Mobile} = 2.87 versus $M_{Desktop}$ = 6.31, p < .001), (ii) less than 20 'Like' votes (M_{Mobile} = 0.21 versus $M_{Desktop}$ = 0.70, p < .001), and (iii) both one or more and less than 20 'Like' votes (M_{Mobile} = 2.21 versus $M_{Desktop}$ = 3.67, p < .001). The same finding is observed when controlling for both the number of photos included in the review post (as *log(Photo+1)*) and the assigned rating (treated a nominal variable) (b = -0.503, se = 0.016, z = -31.257, p < .001; see **Model A2**). Further, the observed negative effect of mobile on review favorability is robust when operationalizing review favorability in terms of number of comments received by the review (b = -0.420, se = 0.023, z = -18.427, p < .001; see **Model A3**).

Table 7. Study 1 (Qunar): The Effects of Mobile on Review Favorability.

	Dependent variables:							
	Likes		Replies			Likes		
	Model A1 (Base Model)	**Model A2** (A1 + Photo and Rating controls)	**Model A3** (A1 but with Replies as DV)	**Model A4** (A1 + Length)	**Model A5** (A1 + Mobile × Expertise interaction)	**Model A6** (A1 but only Novices: Expertise Levels 1, 2)	**Model A7** (A1 but only Experts: Expertise Levels 6, 7)	**Model A8** (A5 + Length)
Mobile	**-0.695*****	**-0.503*****	**-0.420*****	**0.132*****	-0.162***	**-0.274*****	**-1.106*****	0.434***
	(0.016)	**(0.016)**	**(0.023)**	**(0.016)**	(0.021)	**(0.020)**	**(0.060)**	(0.019)
ln(*Length*+1)				0.854***				1.211***
				(0.006)				(0.005)
ln(*ExpertiseLevel*)					1.178***			0.094***
					(0.014)			(0.013)
Mobile × ln(*ExpertiseLevel*)					**-0.688*****			**0.080*****
					(0.019)			**(0.017)**
Constant	-1.560***	0.797***	-0.256***	-3.567***	-2.108***	-2.039***	0.211	-7.588***
	(0.063)	(0.064)	(0.074)	(0.071)	(0.064)	(0.069)	(0.255)	(0.068)
Observations	123,529	123,529	123,529	123,529	123,529	105,506	1,251	123,529
Log Likelihood	-129,716.500	-88,607.110	-58,544.110	-75,316.260	-125,749.500	-93,097.060	-5,433.623	-94,318.450
Akaike Inf. Crit.	259,567.000	177,358.200	117,232.200	150,778.500	251,637.000	186,328.100	10,999.250	188,776.900

Note: *p<0.1; **p<0.05; ***p<0.01

All the models include the following controls: hotel ID, date of review post, and purpose of travel.
Models A1, A2, and A3 show a robust negative effect of mobile on review favorability on the Qunar platform.
Model A4 shows that the negative effect of mobile, as observed in Models A1, A2, and A3, is driven by differences in review length on desktop and mobile reviews.
Model A5, A6 and A7 shows that the overall Qunar-specific negative effect of mobile on review favorability is asymmetric for expert and novice reviewers, where the negative effect is greater for experts.
Model A8 shows that the negative mobile x expertise interaction is driven by differences in review length reductions for expert and novice reviewers.

Consistent with recent published work on mobile-generated reviews (Melumad et al. 2019; Ransbotham et al. 2019), on the Qunar platform, I observe that mobile reviews are significantly shorter in length than desktop reviews ($M_{Desktop}$ = 271.5 versus M_{Mobile} = 61.6 average Chinese characters per review; t(9803) = -40.26, p < .001); note that this is a 77.3% reduction in length from desktop to mobile reviews. Results are robust even when removing extreme values, analyzing only reviews with lengths between the 10th and the 90th percentiles ($M_{Desktop}$ = 69.0 versus M_{Mobile} = 46.5; t(6697) = -54.049, p < .001).

Interestingly, when review length is added as a predictor variable to the base model (**Model A1**), I find that the effect of mobile on review favorability reverses from negative to positive (**Model A4**) (b = 0.132, se = 0.017, z = 8.025, p < .001). Further, running a mediation analysis (Tingley et al. 2013), I find that the effect of generating reviews on mobile (vs. desktop) devices is largely driven by the differences in review length between mobile and desktop reviews on the platform (b = -0.0820; 95% CI: -0.0857, -0.0783; prop. mediation = 73.29%).

My results suggest that whether mobile (vs. desktop) reviews are deemed more or less favorable might actually depend on the level of reduction in review length from desktop to mobile, which may vary across review platforms. I test this idea directly by revisiting published papers on the favorability of mobile reviews across various review platforms and assessing the relationship between degree of reduction in review content from desktop to mobile reviews and the overall conclusion drawn about the favorability of mobile (vs. desktop) reviews.

Using review data from Urbanspoon.com, Ransbotham et al. (2019) find an overall *negative* effect of mobile (vs. desktop) reviews. Interestingly, on Urbanspoon, like Qunar, a very large reduction in review length from desktop to mobile reviews is observed ($M_{Desktop}$ = 81 versus M_{Mobile} = 32); a reduction of 60.5%. Using review data from TripAdvisor.com, Burtch and Hong

(2014) find an overall *positive* effect of mobile (vs. desktop) reviews. In the TripAdvisor, unlike Qunar, a very small reduction in review length from desktop to mobile reviews is observed ($M_{Desktop}$ = 90.49 versus M_{Mobile} = 78.26); a reduction of 13.5%. (I observe a similar pattern of finding in my TripAdvisor dataset in Study 2.) These results are consistent with Mudambi and Schuff (2010) who argue that review length enhances the perceived diagnostic value of reviews.

Therefore, I conclude that whether mobile (vs. desktop) reviews are deemed more or less favorable depends on the level of reduction in review length from desktop to mobile, which varies across review platforms. Analyzing reviews from a platform with a relatively large, e.g., 60%, (small, e.g, 10%) reduction in review length will likely yield an overall negative (positive) effect.

(ii) Distinguishing types of mobile interfaces. Given that the general level of reduction of review content from desktop to mobile reviews on the review platform is a key predictor of the favorability of mobile (vs. desktop) reviews, this begs the question, why do review platforms vary in the degree to which review length is reduced from desktop to mobile?

At this point in time, I do not have the cross-platform type of data that may be required to provide a relatively conclusive answer to such a question. However, I speculate that such differences across platforms likely has to do with differences in *the design of the mobile software interface*. To attempt to illustrate this point empirically, I take advantage of Qunar's distinction of different mobile reviews, reviews generated on (i) their mobile application, (ii) their mobile website, and (iii) SMS texting – where the interface design of the mobile application and the mobile website is a lot richer in content than that of SMS texting. I hypothesize that reviews generated on a mobile interface that is relatively plain in design, lacking cues for review elaboration, such as the case with SMS texting (vs. mobile application and mobile website), are a

lot shorter in review length, which provides less information diagnosticity (Mudambi and Schuff 2010), and therefore would be judged less favorable by readers.

Comparing across these three mobile interfaces on the Qunar platform, as expected, I observe that mobile reviews are not all the same. For example, reviews generated on the mobile application, mobile website and SMS texting, on average, contain 75.6, 70.3, and 33.5 Chinese characters in length per review post, respectively. Coinciding with the theorizing that review length in part affects review favorability by readers (Mudambi and Schuff 2010), for reviews generated on the mobile application and the mobile website, which have very similar review lengths, I observe no significant differences in their review favorability (M_{App} = 0.353 and M_{Site} = 0.373 average 'Like' votes received, *ns*). However, reviews generated on these two interfaces are significantly more favorable than reviews generated on SMS texting (M_{App_Site} = 0.357 versus M_{SMS} = 0.126, $p < .001$). Therefore, I postulate that differences in mobile software interfaces, particularly with regard to how likely the interface engages reviewers to write more content in their reviews (e.g., plain vs. informative background design), is likely a key proximal cause for driving the observed positive/negative effect of mobile (vs. desktop) reviews on judgments of review favorability by readers across review platforms.

*(ii) Mobile * Expertise → Favorability*. Next, I test how the effect of mobile on review favorability might vary as a function of reviewer expertise. Results from my Poisson regression model (see **Model A5** in **Table 7**) shows that the effect of mobile on review favorability is *not* consistent between experts and novice reviewers. Specifically, I find a significant negative interaction between mobile and expertise on review favorability (b = -0.689, se = 0.019, z = -37.144, $p < .001$), where the Qunar-specific negative effect of mobile on review favorability is stronger for experts (*Expertise Levels* 5, 6, and 7: b = -1.106, se = 0.060, z = -18.289, $p < .001$;

see **Model A7**) than novices (*Expertise Levels* 1 and 2: b = -0.274, se = 0.020, z = -13.529, p < .001; see **Model A6**).

I observe that the asymmetric effect of mobile on review favorability for expert and novice reviewers (**Model A5**) is driven by differences in review length (b = 0.080, se = 0.017, z = 4.614, p < .001, see **Model A8**). There is a significant negative interaction between mobile and expertise on review length (b = -0.275, se = 0.020, t(123460) = -14.007, p < .001), where expert reviewers write significantly shorter reviews on mobile (vs. desktop) devices (*Expertise Levels* 5, 6, and 7: b = -1.824, se = 0.106, t(1185) = -17.233, p < .001) than their novice counterparts (*Expertise Levels* 1 and 2: b = -0.854, se = 0.016, t(105439) = -54.106, p < .001).

Conclusions. Based on the results from analyzing Qunar hotel review data, I draw three main conclusions. First, although I find an overall negative effect of mobile on review favorability on the Qunar review platform, I do *not* generalize this finding to all mobile reviews. Drawing on different review platforms (TripAdvisor and Urbanspoon) from past research on the favorability of mobile reviews (Burtch and Hong 2014, Ransbotham et al. 2019, respectively), I conclude that whether mobile (vs. desktop) reviews are deemed more or less favorable by readers depends on the level of reduction in review length from desktop to mobile, which varies across review platforms. For example, for review platforms with a very large reduction in review length from desktop to mobile, such as Qunar and Urbanspoon, with 77.3% and 60.5% reductions, respectively, there is an overall negative effect of mobile on favorability. In contrast, for review platforms with only slight reduction in review length from desktop to mobile, such as TripAdvsior, with a 13.5% reduction, there is an overall positive effect of mobile on favorability.

Second, I show that not all mobile reviews, even on the same platform, are the same. Reviews generated via SMS texting, which is relatively plain in design, compared to the mobile

application and mobile website, are shorter and are deemed less favorable by readers. I postulate that differences in mobile software interfaces is likely one possible proximal cause for driving the observed positive/negative effect of mobile (vs. desktop) reviews on judgments of review favorability by readers across review platforms. However, I believe this topic is something that still needs to be addressed in future research with cross-platform data.

Third, I find that, on Qunar, the overall effect of mobile on review favorability is not consistent across reviewers. Specifically, experts (vs. novices) appear to be particularly hindered in terms of how readers judge their mobile (vs. desktop) reviews. This effect is driven by differences in reduction of review length on mobile (vs. desktop) reviews for expert and novice reviewers.

Although this study, along with past research (Burtch and Hong 2014; Melumad et al. 2019; Ransbotham et al. 2019), shows that the reduction in review content is a consistent feature of generating reviews on mobile (vs. desktop) devices, it is unclear whether the "strategies" adopted by reviewers in writing shorter mobile reviews actually vary. In the following study, I analyze the textual content of TripAdvisor reviews in order to better understand the different approaches adopted by reviewers in writing shorter mobile reviews.

Study 2: TripAdvisor (Field Data)

The purpose of Study 2 is twofold. Firstly, using TripAdvisor reviews, I replicate some of the main findings from Study 1 (Qunar) on readers' favorability of mobile reviews. Secondly, and more importantly, given that the reduction in review length is a consistent feature of generating reviews on mobile (vs. desktop) devices across a number of review platforms (Burtch and Hong 2014; Melumad et al. 2019; Ransbotham et al. 2019), I investigate how the

"strategies" adopted by reviewers in writing shorter mobile reviews vary. Specifically, I examine how the (i) emotionality of language and (ii) density of attributes in mobile (vs. desktop) reviews vary for expert and novice reviewers.

Dataset. For Study 2, I collected and analyzed over 99,000 online reviews over a four year time span, of hotels from TripAdvisor.com, a major online English-based travel review platform (see **Table 4** for description of dataset; see **Table 5** for variable list; see **Table 6** for summary statistics of variables). The dataset only includes reviews posted between January 2012 and December 2016; TripAdvisor's mobile application was first introduced in 2012.

Variables. Similar to Study 1, the main independent variable of interest is *mobile*, which is a binary variable indicating whether the review was generated on a mobile or desktop device. In the dataset, 17.1% of reviews are generated on mobile devices.

The moderating variable of interest is *reviewer expertise*. I operationalize reviewer expertise based on TripAdvisor's platform-defined *0-6 Contributor Level*. Similar to Qunar, TripAdvisor measures their expert reviewers using a points-based system on quality (e.g., inclusion of photos/videos) and quantity of reviews (number of past reviews generated). I used the natural logarithm of TripAdvisor's *0-6 Contributor Level*, i.e., ln(*Contributor_level* + 1), in my analysis to normalize its distribution. Descriptive statistics are provided for reviewers with Contributor Levels less than 2, which make up 60.9% of all reviews in the dataset, and reviewers with *Contributor Levels* greater than 4, which make up 8.6% of all reviews in the dataset.

I investigate the effects of generating reviews on mobile devices for experts and novices across a number of factors including (i) review favorability, (ii) review length, (iii) review emotionality, and (iv) review attribute density. *Review favorability* is operationalized by the number of 'Helpful' votes a review receives. *Review length* is operationalized as the number of

words included in the review, ln(*Length*+1). For robustness of measurement, review length is also operationalized as number of characters in the review.

Review emotionality is the degree of emotional language used in the review. It is calculated using the AFINN word-sentiment dictionary (Hansen et al. 2011). Each word in a review is associated with a specific integer sentiment score, between -5 and 5 (a score of 0 is assigned if the word is not contained in the word-sentiment dictionary). The review emotionality score is calculated by adding the magnitude of the sentiment value of all words in the review divided by the total number of words in the review.

Review attribute density refers to the number of unique attributes included in the review in relation to its length. Review attribute density is calculated by taking the number of unique hotel-related attributes in the review divided by the total number of words in the review, *log(n_attributes/n_words*). Number of attributes is calculated using Part-of-Speech (POS) tagging (Hornik 2016). After POS tagging each word in all hotel reviews in the dataset, I only kept the nouns. Next, I removed city-specific terms by conducting term frequency-inverse document frequency (*tf-idf*) analysis across the six cities. This allowed me to compile 30 of the most frequently used hotel-related nouns; e.g., *service, location,* and *view.* Next, for each review, using a match and count based algorithm, I identified the number of unique nouns mentioned in the review that were contained in the list of 30 hotel-related nouns. This produced my number of unique hotel-specific attributes mentioned in each review. That number was then divided by the total number of words in the review to obtain its review attribute density score.

Analyses. Because my dependent variable of interest, review favorability, is a count variable, I conduct my main analyses using a Poisson regression model. For when my dependent variables are review length, review emotionality, and review attribute density, I use OLS regression.

Included in the analyses are a number of control variables, including hotel ID (*Hotel*), date of review post (converted to number of months from date of review scraping, *MonthsAgo*), and purpose of travel (transformed to five dummy variables, *Purpose*).

$$DV = \beta_0 + \beta_1\, Mobile + \beta_2 \ln(ExpertiseLevel) + \beta_3\, Mobile \times \ln(ExpertiseLevel) + \beta_5\, MonthsAgo + \beta_{6\text{-}10}\, Purpose + \beta_{11\text{-}70}\, Hotel + \varepsilon$$

Results: (i) Mobile → Favorability. Consistent with results from Study 1 and recent published work on mobile-generated reviews (Melumad et al. 2019; Ransbotham et al. 2019), on the TripAdvisor platform, I find that mobile reviews are significantly shorter in length than desktop reviews (in terms of characters per review: $M_{Desktop} = 617$ vs. $M_{Mobile} = 551$; $t(31885) = 17.863$, $p < .001$; in terms of number of words per review: $M_{Desktop} = 113$ vs. $M_{Mobile} = 101$; $t(31691) = 17.292$, $p < .001$). However, unlike my Qunar review data where I observed a very large reduction, 77.3%, in length from desktop to mobile reviews, in my TripAdvisor review data, I observe only a slight reduction, 10.0%.

Consistent with my theorizing from Study 1 on how the favorability of mobile reviews varies across platforms, with a relatively small reduction in review length on the TripAdvisor platform, I find an overall *positive* effect of generating reviews on mobile (vs. desktop) devices on judgments of review favorability by readers ($M_{Desktop} = 0.578$ vs. $M_{Mobile} = 0.631$ average 'Helpful' votes per review; $b = 0.058$, $se = 0.011$, $z = 5.150$, $p < .001$; see **Model B1** in **Table 8**). This finding is consistent with Burtch and Hong (2014) who also analyzed TripAdvisor review data and find a positive effect of mobile on review favorability. The finding is robust when also controlling for both the number of photos included in the review post (as *log(Photo+1)*) and the

Table 8. Study 2 (TripAdvisor): The Effects of Mobile on Review Favorability.

	Dependent variable:						
	Helpful						
	Model B1 (Base Model)	**Model B2** (B1 + Photo and Rating controls)	**Model B3** (B1 + Length)	**Model B4** (B1 + Mobile * Expertise interaction)	**Model B5** (B1 but only Novices: Contributor Levels $\leq$ 2	**Model B6** (B1 but only Experts: Contributor Levels $\geq$ 4	**Model B7** (B5 + Length)
Mobile	**0.058*****	**0.032*****	**0.061*****	0.157***	**0.091*****	**-0.045**	0.135***
	(0.011)	**(0.011)**	**(0.011)**	(0.024)	**(0.016)**	**(0.034)**	(0.024)
ln(*Length*+1)			0.328***				0.337***
			(0.008)				(0.008)
ln(*ExpertiseLevel*+1)				0.049***			-0.025***
				(0.007)			(0.007)
Mobile × ln(*ExpertiseLevel*+ 1)				**-0.084*****			**-0.054*****
				(0.017)			**(0.017)**
Constant	-1.521***	-0.172***	-3.489***	-1.571***	-1.547***	-2.021***	-3.514***
	(0.032)	(0.037)	(0.057)	(0.033)	(0.043)	(0.113)	(0.057)
Observations	99,050	99,050	99,050	99,050	61,085	8,325	99,050
Log Likelihood	-98,173.020	-95,503.520	-97,266.590	-98,144.990	-59,747.860	-7,492.308	-97,249.210
Akaike Inf. Crit.	196,480.100	191,151.000	194,669.200	196,428.000	119,629.700	15,118.620	194,638.400

Note: *p<0.1; **p<0.05; ***p<0.01

All the models include the following controls: hotel ID, date of review post, and purpose of travel.
Models B1, B2, and B3 demonstrate the robustness of the positive effect of mobile on review favorability on the TripAdvisor platform.
Model B4 demonstrates how the effect of mobile on review favorability varies as a function of reviewer expertise.
Models B5 and B6 demonstrate that the positive effect of mobile occurs for novices, but not experts.
Model B7 shows that the negative mobile x expertise interaction, as observed in Model B4, is driven by differences in review length reductions for expert and novice reviewers.

assigned rating (treated a nominal variable) (b = 0.032, se = 0.011, z = 2.774, p = .0055; see **Model B2**).

Similar to Study 1 results, when review length is added as a predictor variable to the base model (Model B1 in Table 5), the effect of mobile on review favorability is positive (b = 0.061, se = 0.011, z = 5.389, p < .001; see **Model B3**). Further, running a mediation analysis (Tingley et al. 2013), I find that the effect of generating reviews on mobile (vs. desktop) devices is in part driven by the differences in review length between mobile and desktop reviews on the platform (b = 0.00227; 95% CI: 0.00157, 00295; prop. mediated = 11.1%).

*(ii) Mobile * Expertise → Favorability*. Next, I test how the effect of generating reviews on mobile devices might vary as a function of reviewer expertise. Results from my Poisson regression model (see **Model B4**) shows that the effect of mobile on review favorability is *not* consistent between expert and novice reviewers. In line with results from Study 1, I find a significant negative interaction between mobile and expertise on review favorability (b = -0.084, se = 0.017, z = -4.925, p < .001), where the TripAdvisor-specific positive effect is salient for novices (*Contributors Levels* < *2*, b = 0.091, se = 0.016, z = 5.495, p < .001; see **Model B5**), but not experts (*Contributor Levels* > 4, b = -0.044, se = 0.033, z = -1.321, *ns*; see **Model B6**).

Similar to Study 1, I find that the asymmetric effect of mobile on review favorability for expert and novice reviewers (Model B4) is driven by differences in review length (b = -0.054, se = 0.017, z = -3.129, p = .002; **Model B7**). There is a significant negative interaction between mobile and expertise on review length (b = -0.236, se = 0.010, t = -23.974, p < .001), where the reduction in review length from desktop to mobile is more pronounced for experts ($M_{Desktop}$ = 137 vs. M_{Mobile} = 107; b = -0.194, se = 0.015, t = -13.15, p < .001) than novices ($M_{Desktop}$ = 106 vs. M_{Mobile} = 98; b = -0.016, se = 0.007, t = -2.173, p = 0.03).

With the reduction in review length being a consistent feature of generating reviews on mobile (vs. desktop) devices across a number of review platforms (Burtch and Hong 2014; Melumad et al. 2019; Ransbotham et al. 2019), in the subsequent sections, I investigate how the "strategies" adopted by expert and novice reviewers in writing shorter mobile reviews might vary. Specifically, I examine how the emotionality of language and density of attributes in mobile (vs. desktop) reviews vary for expert and novice reviewers.

(iii) Review Content: Emotionality. Consistent with past research on mobile reviews (Melumad et al 2019; Ransbotham et al. 2019), in my TripAdvisor review data, I find that mobile (desktop) reviews contain more emotional language (b = 0.0055, se = 0.0007, z = 7.512, p < .001; see **Model C1** in **Table 9**). However, the enhanced use of emotional language on mobile (vs. desktop) reviews is *not* consistent across reviewers. There is a significant negative interaction between mobile and expertise on the emotionality of language used in the reviews (b = -0.0039, se = 0.0014, z = -3.169, p = .0015; see **Model C2**), where the enhanced emotionality of mobile (vs. desktop) reviews is more pronounced for novices (*Expertise Levels* < 2; b = 0.0079, se = 0.0011, z = 6.468, p < .001; see **Model C3**) than experts (*Expertise Levels* >4; b = 0.0056, se = 0.0019, z = 3.031, p = .002; see **Model C4**).

(iii) Review Content: Review Attribute Density. Although no main effect of mobile on attribute density is observed (see **Model 1** in **Table 10**), results from my OLS regression model (**Model D2**) shows that mobile (vs. desktop) reviews vary in their attribute density as a function of reviewer expertise (b = 0.043, se = 0.007, t = 6.096, p < .001), where experts generate mobile (vs. desktop) reviews that are *more* attribute dense (4.9% more dense, b = 0.033, se = 0.011, t = 3.148, p = .002; see **Model D4**), novice generate mobile (vs. desktop) reviews that are *less* attribute dense (1.8% less dense, b = -0.017, se = 0.006, t = -2.712, p = .007; see **Model D3**).

Table 9. Study 2 (TripAdvisor): The Effect of Mobile and Expertise on Review Emotionality.

	Dependent variable:			
	Review Emotionality			
	Model C1 (Base Model)	**Model C2** (+Expertise)	**Model C3** (Novices)	**Model C4** (Experts)
Mobile	**0.0049*****	0.0107***	**0.0079*****	**0.0056*****
	(0.0008)	(0.0017)	**(0.0015)**	**(0.0019)**
ln(*ExpertiseLevel*+1)		-0.0045***		
		(0.0006)		
Mobile × ln(*ExpertiseLevel*+ 1)		**-0.0039*****		
		(0.0014)		
Constant	0.6189***	0.6123***	0.6384***	0.5690***
	(0.0040)	(0.0041)	(0.0059)	(0.0154)
Observations	98,626	98,626	48,559	8,279
R^2	0.2428	0.2436	0.2362	0.2071
Adjusted R^2	0.2427	0.2435	0.2361	0.2065
Residual Std. Error	0.0900 (df = 98619)	0.0900 (df = 98617)	0.0968 (df = 48552)	0.0796 (df = 8272)
F Statistic	5,269.2260*** (df = 6; 98619)	3,969.3180*** (df = 8; 98617)	2,502.3050*** (df = 6; 48552)	360.0743*** (df = 6; 8272)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 10. Study 2 (TripAdvisor): The Effect of Mobile and Expertise on Review Attribute Density.

	Dependent variable:			
	Review Attribute Density [log(n_attributes/n_words)]			
	Model D1 (Base Model)	**Model D2** (+Expertise)	**Model D3** (Novices)	**Model D4** (Experts)
Mobile	**-0.005**	-0.040***	**-0.017*** **	**0.033*** **
	(0.004)	(0.009)	**(0.006)**	**(0.011)**
ln(*ExpertiseLevel*+1)		-0.058***		
		(0.003)		
Mobile × ln(*ExpertiseLevel*+1)		**0.043*** **		
		(0.007)		
Constant	-2.697***	-2.638***	-2.651***	-2.774***
	(0.008)	(0.009)	(0.011)	(0.025)
Observations	98,518	98,518	60,699	8,284
R^2	0.020	0.024	0.024	0.030
Adjusted R^2	0.020	0.024	0.023	0.022
Residual Std. Error	0.452 (df = 98451)	0.451 (df = 98449)	0.461 (df = 60632)	0.442 (df = 8217)
F Statistic	30.801*** (df = 66; 98451)	36.008*** (df = 68; 98449)	22.437*** (df = 66; 60632)	3.873*** (df = 66; 8217)
Note:				*p<0.1; **p<0.05; ***p<0.01

This finding is consistent with Schema Theory (Axelrod 1973; Mandler 2014), where because of their review-writing experience, expert reviewers are expected to be relatively consistent in their topic coverage on mobile and desktop reviews. Even though generating reviews on mobile devices are shorter, experts appear to "compensate" by generating reviews that are *more* attribute dense. In contrast, novices write mobile (vs. desktop) reviews that contain proportionately *less* attributes, devoting more attention on elaborating on attended attributes.

Next, I test the interaction between review attribute density and review length on readers' favorability judgments of reviews. First, I find significant positive main effects for review attribute density ($b = 7.433$, $se = 1.004$, $z = 7.403$, $p < .001$) and review length on review favorability ($b = 0.451$, $se = 0.019$, $z = 23.645$, $p < .001$). I find their interaction to be negative ($b = -1.892$, $se = 0.239$, $z = -7.932$, $p < .001$), demonstrating that the positive effect of attribute density on readers' favorability judgments is greater for shorter reviews. This finding is in line with the theory of information overload (Jacoby 1974, 1984) and research emphasizing the consequences of reviews that provide too much information (Park and Lee 2008).

Conclusion. Three main conclusions are drawn from Study 2. First, although I find a positive effect of mobile on review favorability on the TripAdvisor platform, I do not generalize this finding to all mobile reviews. Instead, I draw the conclusion that the relationship between mobile and review favorability depends on the general level of reduction in review length from desktop to mobile reviews.

Second, I find differences in the "strategies" adopted by expert and novice reviewers in generating shorter mobile reviews – expert reviewers are more consistent in generating reviews on both device types. Although mobile (vs. desktop) reviews have been found to be more emotional in content (Melumad et al. 2019; Ransbotham et al. 2019), this effect is mitigated for

experts relative to novices. Although mobile (vs. desktop) reviews are known to be shorter (Melumad et al. 2019; Ransbotham et al. 2019), experts generate mobile reviews that are more attribute dense (4.9% more dense), whereas novices generate mobile reviews that are less attribute dense (1.8% less dense). This finding suggests that experts compensate in their mobile reviews by including proportionately more attributes.

Finally, I find that although review attribute density has a positive effect on readers' favorability judgments of reviews, this effect is particularly pronounced for shorter reviewers.

General Discussion

In this essay, I examined how generating mobile (vs. desktop) reviews vary across (i) review platforms and (ii) reviewer expertise. I find that whether mobile (vs. desktop) reviews are deemed more or less favorable by readers largely depends on on the general level of reduction in review length from desktop to mobile reviews, where platforms with a large (small) reduction likely yield an overall negative (positive) effect of mobile. I postulate that differences in mobile software interfaces is likely one proximal cause for observing conflicting findings on readers' favorability judgments of mobile (vs. desktop) reviews across review platforms (Burtch and Hong 2014; Ransbotham et al. 2019).

Consistent with Schema Theory (Axelrod 1973; Mandler 2014), where expert reviewers are expected to be relatively consistent in their reviews regardless of device type, I find that the enhanced use of emotional language in mobile (vs. desktop) reviews is more pronounced for novices than experts. Although mobile (vs. desktop) reviews are shorter for both experts and novices (Burtch and Hong 2014), I find that experts (novices do not) "compensate" by

generating mobile reviews that are more (less) attribute dense. Interestingly, I find that this more-attribute-dense mobile strategy by experts is particularly effective for shorter reviews.

This research provides two important practical implications. First, for a long time, the feature to generate reviews on mobile devices has largely been avoided by review platforms in fear that users would write reviews in an irrational or emotional state. For example, prior to 2013, Yelp only had a "Quick Tips and Draft Reviews" mobile feature that provided eager Yelp reviewers with an outlet to jot notes about their immediate experiences that they can then add to or edit later when they got back to a desktop computer[1]. Although my research does find a fair degree of emotional language used in mobile reviews, this enhanced emotionality in mobile reviews is very much attenuated for expert reviewers. This finding suggests that any measure taken by review platforms to avoid users generating reviews on mobile reviews should not be applied to all users, but rather narrowed to only novice users.

Second, a major goal for both service providers and review platforms is for past customers to not only write reviews, but also provide a fair amount of detail about their customer experience, especially when the experience is very good. In turn, this information can help prospective review-reading customers make their consumption choice. Given that increasingly more reviews are generated on mobile devices (Yelp 2019), the major issue I find in my research is that there is a considerable reduction in length for reviews generated on mobile (vs. desktop) devices, thus, limiting the diagnostic value that mobile reviews provide to prospective customers (Mudambi and Schuff 2010). To be fair, the degree of reduction does vary quite substantially across platforms. For example, Qunar and Urbanspoon has a large reduction of 77.3% and 60.5%, respectively, whereas TripAdvisor has a slight reduction of 10%. Looking into the

[1] https://blog.yelp.com/2009/12/ask-yelp-why-cant-i-write-reviews-from-my-mobile

mobile application of TripAdvisor, I speculate that the informative background design on the mobile interface likely plays a role in reviewers elaborating in their mobile reviews. For example, on their mobile application, TripAdvisor has reviewers not only provide an overall star rating and write a review, but also consider the experience across a number of experience-related dimensions (e.g., value, location, service quality). I postulate that such an informative background design can act as cues for reviewers to elaborate about their experiences on mobile devices.

This research has a few important limitations. First, because my datasets in this research consisted of reviews on specific service providers, rather than reviews by specific reviewers, I can extend my findings to between-reviewers, but not within-reviewers. Thus, my results are susceptible to the possibility of *self-selection biases* driving some of the observed effects. I believe this is the main limitation to the existing version of this essay. However, the pattern of results in my research are consistent with past research papers on mobile reviews that collected and analyzed reviews by a number of reviewers (Burtch and Hong 2014; Ransbotham et al. 2019). Further, I replicate many of my results across two different review platforms: Qunar, a major travel review platform in Chinese, and TripAdvisor, a major travel review platform in English. Although I do not speculate that my results are driven by self-selection biases, to ensure robustness of findings, it would be best to collect an additional set of reviews from a number of *reviewers*, instead of service providers, and test whether the results are replicated. This would allow me to clearly rule out concerns about self-selection bias. Additionally, running randomized controlled experiments can help mitigate concerns of self-selection bias, as well as strengthen claims of causality.

Second, most of my analyses focused on main effects and interactions, with limited attention to mechanisms. Given the two key features of mobile devices – their portability and their constraining interfaces (Burtch and Hong 2014) – it is unclear the extent to which the observed effects in my research are driven by each of these mobile device features. Past research on mobile reviews would suggest that many of the observed effects, such as the enhanced emotionality of mobile reviews, are multiply-determined, where both the portability and the constraining interface features drive the effect (Burtch and Hong 2014; Melumad et al. 2019). A combination of collecting more fine-tune time-stamped data and running a series of experiments would help parse out the extent to which observed findings may be driven by each mobile feature.

Third, a central theme in this essay is the importance of drawing conclusions based on findings across more than one platform. For example, my conclusion drawn on the favorability of mobile reviews were based on my review data from two platforms, as well as review data from published papers on mobile reviews (Burtch and Hong 2014; Ransbotham et al. 2019). However, my conclusions drawn about differences in review content as a function of device type and reviewer expertise was only based on a single review platform, TripAdvisor. Collecting additional data, in particular, reviews from a number of reviewers, would strengthen the generalizability of my results.

An important notion alluded in this research is that mobile is not purely binary. In this research, I demonstrate that mobile reviews vary (i) across review platforms, (ii) within review platform, and (iii) across reviewer expertise levels. Although extant research has focused on comparing and contrasting reviews generated on mobile (vs. desktop) devices (e.g., Melumad et al. 2019; Ransbothem et al. 2019), I believe that future research should begin to embrace the

nuances of mobile, studying why and how mobile reviews vary. As emphasized in this research, I speculate that the observed effects of mobile are driven by not only the device, but also the software interface. Future research can go beyond studying mobile devices and explore how various aspects of the mobile interface design shape how consumers generate their review content.

Conclusions. Given the ubiquity of mobile devices in the hands of consumers and the increasing prevalence of mobile-generated reviews, this research demonstrates the effects of generating reviews on mobile (vs. desktop) devices and how the expertise of consumers in generating reviews plays an important role on the effects of mobile. As new technologies emerge and become mainstream, the topic of how technological mediums shape the way consumers generate, share, and consume content will continue to be important to the area of consumer research.

References

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

Ariely, Dan, and George Loewenstein. "The heat of the moment: The effect of sexual arousal on sexual decision making." *Journal of Behavioral Decision Making* 19, no. 2 (2006): 87-98.

Axelrod, Robert (1973), "Schema Theory: An Information Processing Model of Perception and Cognition." *American Political Science Review*, 67 (4), 1248-66.

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.

Beilock, Sian L., and Thomas H. Carr (2001), "On the Fragility of Skilled Performance: What Governs Choking Under Pressure?" *Journal of Experimental Psychology: General*, 130 (4), 701-25.

Burtch, Gordon, and Yili Hong (2014), "What Happens When Word of Mouth goes Mobile?" *Thirty Fifth International Conference on Information Systems, Auckland*.

Business Insider Intelligence (2016), "Mobile Apps are still Dominating Users' Time," http://www.businessinsider.com/mobile-apps-are-still-dominating-users-time-2016-9

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.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston (2008), "Do Online Reviews Matter? An Empirical Investigation of Panel Data," *Decision Support Systems*, 45 (4), 1007-16.

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.

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

Ghose, Anindya, and Panagiotis G. Ipeirotis (2011), "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics." *IEEE Transactions on Knowledge and Data Engineering*, 23 (10), 1498-512.

Goldstein, Alvin G., and June E. Chance (1980), "Memory for Faces and Schema Theory," *The Journal of Psychology*, 105 (1), 47-59.

Grewal, Lauren, and Andrew T. Stephen (2019), "In Mobile We Trust: The Effects of Mobile Versus Nonmobile Reviews on Consumer Purchase Intentions," *Journal of Marketing Research*, 1-18.

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.

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

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.

Jacoby, Jacob (1974), "Consumer Reaction to Information Dis- plays: Packaging and Advertising," in Advertising and the Public Interest, ed. S. F. Divita, Chicago: American Mar- keting Association, 101-118.

Jacoby, Jacob. "Perspectives on information overload." *Journal of consumer research* 10, no. 4 (1984): 432-435.

Johnson, E., and Payne, J. 1985. "Effort and Accuracy in Choice," *Management Science*, 31 (4), 395-415.

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.

Kusumasondjaja, Sony, Tekle Shanka, and Christopher Marchegiani (2012), "Credibility of Online Reviews and Initial Trust: The Roles of Reviewer's Identity and Review Valence," *Journal of Vacation Marketing*, 18 (3), 185-95.

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

Loewenstein, G. (1996), "Out of Control: Visceral Influences on Behavior," *Organizational Behavior and Human Decision Processes*, 65 (3), 272-92.

Loewenstein, G. 2000. "Emotions in Economic Theory and Economic Behavior," *The American Economic Review*, 90 (2), 426-32.

Ludwig, Stephan, Ko De Ruyter, Mike Friedman, Elisabeth C. Brüggen, Martin Wetzels, and Gerard Pfann (2013), "More than Words: The Influence of Affective Content and

Linguistic Style Matches in Online Reviews on Conversion Rates," *Journal of Marketing*, 77 (1), 87-103.

Mandler, Jean Matter. *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Psychology Press, 2014.

Melumad, Shiri, J. Jeffrey Inman, and Michel Tuan Pham (2019), "Selectively Emotional: How Smartphone Use Changes User-Generated Content," *Journal of Marketing Research*, 56 (2), 259-75.

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

Moore, Sarah G. (2015), "Attitude Predictability and Helpfulness in Online Reviews: The Role of Explained Actions and Reactions," *Journal of Consumer Research*, 42 (1), 30-44.

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.

Nielsen 2016, “Millenials are Top Smartphone Users,” http://www.nielsen.com/us/en/insights/news/2016/millennials-are-top-smartphone-users.html

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.

Ransbotham, Sam, Nicholas H. Lurie, and Hongju Liu (2019), "Creation and Consumption of Mobile Word of Mouth: How Are Mobile Reviews Different?" *Marketing Science*, 1-20.

Rentsch, Joan R., Tonia S. Heffner, and Lorraine T. Duffy (1994), "What You Know is What You Get from Experience: Team Experience Related to Teamwork Schemas," *Group & Organization Management*, 19 (4), 450-74.

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."

Yelp 2019, "An Introduction to Yelp Metrics as of March 31, 2019," https://www.yelp.ca/factsheet

Yin, Dezhi, Samuel Bond, and Han Zhang (2013), "Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews," *MIS Quarterly*, 38 (2), 539-60.

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.

Ziefle, Martina (2002), "The Influence of User Expertise and Phone Complexity on Performance, Ease of Use and Learnability of Different Mobile Phones," *Behaviour & Information* Technology, 21 (5), 303-11.

Final Remarks

In this dissertation, I advance our understanding about the shared content generated by consumers on online review platforms, specifically, disentangling the role of reviewer expertise and mobile devices in the generating of shared review content. Looking beyond the research findings of Essays 1 and 2, I think there are three important ideas to take away, particularly as I move through the next stages of my research career. These ideas include: (i) the importance of aggregate (summary) metrics in an age where user-generated content is becoming increasingly abundant, (ii) the contribution to theory in a way that extends across multiple platforms, and (iii) the value of combining real world data with behavioral experiments in order to enhance the external and internal validity of consumer research.

One of the major trends that we are observing in the online space is that increasingly more content (e.g., reviews, blogs, and videos) is being produced by consumers. With the flourishing of user-generated content, consumers are less likely to access and consume all the available content. Instead, they rely more and more on aggregate-level measures that summarize the abundance of content in order to guide their consumption choice. Although much of the published research on online reviews examines reviews either at the aggregate (Babić Rosario et al. 2016; Chevalier and Mayzlin 2006) or individual level (Melumad, Inman, and Pham 2019; Packard and Berger 2017), little to no research has explored the interaction between the two levels. Consider the consumers' navigation process on online review platforms. In many instances, consumers navigate back and forth between aggregate and individual review levels, where they might use aggregate metrics, like user rating averages and number of reviews, to guide which restaurants to consider, and then read individual reviews about the selected restaurants to help make their choice. Capturing the interdependencies between aggregate and

individual review levels will become increasingly important with the abundance of openly available user-generated content.

A central theme in this dissertation is the importance of drawing conclusions that are based on findings from multiple platforms, or datasets. A major goal for us researchers is to contribute to theory in a way that is generalizable across people, place, and time. I think that finding effects in a single dataset and connecting them with theory is a reasonable approach in science. However, in the age of "big data" where the number of observations in a dataset can be in the hundreds of thousands, if not millions, we, as researchers, need to be concerned about whether the observed effects are by chance or are truly meaningful and reflective of the real world. Using theory to help explain an observed finding is important, however, given that the scientific literature is quite expansive where theories for A and $\bar{A}$ likely both exist, I believe that a more robust approach would be to not only tie results to theory, but also replicate the findings across different platforms and reconcile any cross-platform differences. Therefore, as I move forward in my research career, I believe that placing emphasis on the replicability across platforms is a valuable compass to my research endeavors.

Finally, as consumers generate increasingly more content and the ability to collect that content becomes increasingly accessible, I believe that the field of consumer research will place greater value in studying consumer-relevant phenomenon from both real world data (e.g., online reviews) and behavioural experiments (Inman et al. 2018). Where the value of observing a phenomenon with real world data is in its generalizability, the benefit of establishing a phenomenon with randomized control experiments is in drawing claims about causality. Therefore, as I move through the next stages of my research career, I believe that using a mixed-method approach will be central to my research methodology.

To conclude, the goal of this dissertation was to advance our collective understanding about consumer-generated review content. However, the journey in achieving this goal has empowered me with a number of research tools, theoretical and methodological, and has contributed to my excitement in continuing to conduct research on many of the emerging topics in the area of technology and consumer behaviour.

References

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.

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.

Inman, J. Jeffrey, Margaret C. Campbell, Amna Kirmani, and Linda L. Price (2018), "Our Vision for the Journal of Consumer Research: It's All About the Consumer," *Journal of Consumer Research*, 955-59.

Melumad, Shiri, J. Jeffrey Inman, and Michel Tuan Pham (2019), "Selectively Emotional: How Smartphone Use Changes User-Generated Content," *Journal of Marketing Research*, 56 (2), 259-75.

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

Appendix A

Stimuli for Study 2.

Please think about some of the *restaurants you have visited* [*electronics products you have purchased*] over the past year.

Please list the name of three of these *restaurants* [*electronic products*] and rate your experience with each.

	Terrible	**Poor**	**Average**	**Very Good**	**Excellent**
1. _____	1	2	3	4	5
2. _____	1	2	3	4	5
3. _____	1	2	3	4	5

Imagine you have just eaten at the "Amsterdamn BrewHouse - on the Lake", a restaurant located in Toronto, and have written the following review:

Positive Condition

I really loved the atmosphere at this place. It's rustically modern interior design is great - it has high ceiling, wooden tables, and large windows. What fascinated me was that even though we were there around noon, the place was very busy; however, it didn't feel overcrowded at all. Service was friendly and helpful, and our food was tasty. They've got quite a selection of beers, local as well as international. Being tourists, we ordered two different Canadian brands.

Negative Condition

The restaurant is located by the lake, but the great view cannot offset the bad service and food. Trying to get a seat on the front patio was near impossible -- there was a 45 min wait on a Wednesday at 2pm. The service was not so great. When I asked for vegetarian options, the waiter was clueless. Veggieburger is rice formed into a patty. Portions were small and overpriced.

How would you rate this restaurant?

1- Terrible 2- Poor 3 - Average 4 - Very Good 5 - Excellent

Appendix B

Sample of Stimuli for Study 2B

Please take a moment to recall a relatively recent positive experience at a sit-down restaurant.

What is the name of this restaurant?

How many weeks ago (approximately) did you visit this restaurant?

How many times in total have you visited this restaurant?

- 1 time
- 2-5 times
- 6-10 times
- More than 10 times

Please rate your restaurant experience based on the following dimensions:

	Terrible 1 Star	Poor 2 Stars	Average 3 Stars	Very Good 4 Stars	Excellent 5 Stars
Location					
Food					
Atmosphere					
Cleanliness					
Service					
Value					

Please rate your overall restaurant experience:

Terrible 1 Star	Poor 2 Stars	Average 3 Stars	Very Good 4 Stars	Excellent 5 Stars

Curriculum Vita

Name: Peter Nguyen

Education: Doctor of Philosophy – Ivey Business School, Western University (2019)
Marketing

Master of Science – Western University (2013)
Neuroscience

Bachelor of Science – Western University (2011)
Major in Psychology
Major in Medical Sciences

Awards: Joseph-Armand Bombardier Canada Graduate Scholarship 2014-2017, $105k

Ontario Graduate Scholarship (OGS) 2014, $15k

Teaching Experience: Assistant Professor of Marketing
Miami University
2019-

Instructor
Western University
2018-2019

Research Papers:

Nguyen, Peter, Xin (Shane) Wang, Xi Li, & June Cotte. *Expert reviewers' restraint from extremes and its impact on service providers*. (Under 2nd round review at ***Journal of Consumer Research***)

Nguyen, Peter, Xin (Shane) Wang, & David J. Curry. *Unraveling postulated psychological explanations for the endowment effect: A meta-analysis*. (Invited for 2nd round review at ***Organizational Behavior and Human Decision Processes***)