

2021

Empirical studies of online markets: the impact of product page cues on consumer decisions

<https://hdl.handle.net/2144/42547>

Downloaded from OpenBU. Boston University's institutional repository.

BOSTON UNIVERSITY
QUESTROM SCHOOL OF BUSINESS

Dissertation

**EMPIRICAL STUDIES OF ONLINE MARKETS: THE
IMPACT OF PRODUCT PAGE CUES ON CONSUMER
DECISIONS**

by

SHRABASTEE BANERJEE

B.S., Lady Brabourne College, Calcutta University, 2013
M.S., University of Warwick, 2014

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2021

© 2021 by
SHRABASTEE BANERJEE
All rights reserved

Approved by

First Reader

Georgios Zervas, PhD
Associate Professor of Marketing

Second Reader

Anita Rao, PhD
Associate Professor of Marketing
Booth School of Business
University of Chicago

Third Reader

Chrysanthos Dellarocas, PhD
Associate Provost for Digital Learning and Innovation
Richard C. Shipley Professor of Management

**Don't Panic.*

**I love deadlines. I like the whooshing sound they make as they fly by.*

**A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools.*

**I may not have gone where I intended to go, but I think I have ended up where I needed to be.*

-Douglas Adams

Acknowledgments

All PhD journeys are long, arduous and rewarding in varying measures, with the mix depending heavily on your luck and your support systems. I was fortunate to have a (largely!) rewarding PhD journey and this would not have been possible without those I am about to thank here. Firstly, thanks to my advisor Georgios for helping me become an independent researcher and giving me the support needed to get where I wanted to be. When I started grad school, I had very limited experience with coding and the specifics of academic marketing research. I will always be grateful to Georgios for his patience in those initial days, and for giving me the time I needed to learn things even if it was less than optimal. I am also grateful for his support and advice as I navigated the job market and negotiated offers. Next, thanks to my wonderful committee members Chris and Anita, who, despite their busy schedules, have always found time to give me feedback and enrich my projects. I hope I can continue to learn from them and grow as I find my way in this profession.

I also want to thank the Questrom School of Business for providing generous conference funding that allowed me to build my academic network and interact with the great community of scholars working in my interest areas (and beyond). Thanks also to the Hariri Institute of Computing for choosing me as a Graduate Student Fellow and supplementing my research expenses through their fellowship.

A big part of my PhD experience was fueled by the advantages of being in Boston. To that end, I want to thank the many Boston-area schools for providing an environment of openness and collaboration. This environment definitely extends beyond just academics - the past year has been so hard for everyone around the world, and I am so grateful to the wonderful community I have found here (which includes kind neighbors, independent bookstores and local parks) to make the 'lockdown' experience bearable.

I want to thank my friends and family for their constant support: my cohort-mates at Questrom, Mohit and Akshita (my best friends in Boston who are now like family), my friends from back home (Priyanka, Srabana, Sandipa and Suddhasatwa) who are never tired of my cribbing, my cousin Apratim for helping me navigate the American life with his advice and humor, and my parents for their love and encouragement through the years.

Finally, a special shout-out to two very important men in my life. First, my father Somprakash, who was my earliest collaborator and academic mentor (and continues to fulfil that role!) for all his support. Next, my partner Narendra, for seeing me through this PhD, always making time to correct code and proofread documents, keeping me well-fed and well-watered, and for all the adventures we have had together.

There are so many others (my family back home, and colleagues across the profession) who have helped me in big and small ways throughout this journey. Thanks to all of them for making this possible.

EMPIRICAL STUDIES OF ONLINE MARKETS: THE IMPACT OF PRODUCT PAGE CUES ON CONSUMER DECISIONS

SHRABASTEE BANERJEE

Boston University, Questrom School of Business, 2021

Major Professor: Georgios Zervas, PhD
Associate Professor of Marketing

ABSTRACT

The widespread expansion of online markets in the past decade poses several questions for platforms, firms and customers alike. An important dimension to be explored in this domain is the provision of information on e-commerce platforms - given the increasing ease with which product pages can be customized to include a vast variety of content, how do these pieces of information interact? Further, what are the specific channels through which this information eventually influences consumer decision-making? My dissertation is situated in this space, and aims to look at how consumers respond to various “cues” that are being introduced by e-commerce platforms which offer products or services that can be purchased online, and how these cues might eventually influence decision-making. In my first dissertation project, the cue I focus on is user generated content. More specifically, I study how the introduction of the Q&A technology (which enables customers to ask product-specific questions before purchase, and receive answers either from other customers or the platform itself) affects the more widely established reviews and ratings feature on e-commerce platforms. I find that the addition of Q&As leads to better matches

between customers and products, higher customer satisfaction, and resultantly higher ratings. My second project examines another cue that is common in online markets, which is the advertised reference price. My goal in this project is to examine how users react to a specific variant of such prices, namely the “Starting from...” price, using data from a large scale field experiment conducted on Holidu.com. My results indicate that raising “From” prices gives users a more accurate price estimate, but it negatively impacts outbound clicks and other engagement metrics. Taken together, the two projects aim to shed light on factors that influence consumer decision-making in an e-commerce setting, and the possible mechanisms underlying this influence.

Contents

1	Introduction	1
2	Interacting User Generated Content Technologies: How Questions and Answers Affect Consumer Reviews	4
2.1	Introduction	5
2.2	Related work	8
2.3	Conceptual framework	12
2.4	Data and descriptives	14
2.4.1	Q&A, reviews, and fit uncertainty	15
2.5	Empirical Strategy	17
2.5.1	Measuring fit mismatch	20
2.5.2	Mean reversion and measurement error	24
2.6	Results	28
2.6.1	Low pre-Q&A ratings	28
2.6.2	High pre-Q&A variance	29
2.6.3	High proportion of pre-Q&A fit-related negative reviews	30
2.6.4	Mechanism: fit mismatch reduction	30
2.7	Conclusion	36
2.8	Tables	38
2.9	Figures	49
3	Reference Price Effects In Vacation Rental Markets	54
3.1	Introduction	56

3.2	Related work	60
3.2.1	Advertised Reference Prices	60
3.2.2	Price obfuscation and salience	62
3.2.3	Fairness perceptions	64
3.3	Conceptual framework	65
3.4	Experimental details	68
3.4.1	Experiment design	68
3.4.2	Data	69
3.5	Results	71
3.5.1	ATE of high floor prices	71
3.5.2	ATE for users exposed to actual prices	72
3.5.3	Treatment effect heterogeneity: CATE	76
3.6	Conclusion	80
3.7	Tables	83
3.8	Figures	91
4	Conclusion	100
A	Interacting User Generated Content Technologies: How Questions and Answers Affect Consumer Reviews	105
A.1	Tables and Figures	105
A.2	Mathematical Appendix	116
B	Reference Price Effects	125
B.1	Doubly Robust Estimation of Heterogeneous Treatment Effects	125
B.2	CATE estimates: doubly robust	128
B.3	Some correlational evidence from Airbnb	132
B.3.1	Within listings	133
B.3.2	Across listings	138

List of Tables

2.1	Summary statistics.	38
2.2	Top-10 product categories ordered by the average number of questions received.	39
2.3	Top-3 LDA topics for reviews and Q&As.	39
2.4	Top-5 reviews with the highest predicted probabilities of quality and fit issues.	40
2.5	Top-20 words most predictive of quality/fit issues, estimated using layerwise relevance propagation on the output of an SVM classifier.	40
2.6	The impact of Q&A on ratings using low pre-treatment ratings (≤ 4) as a measure for fit mismatch.	41
2.7	The impact of Q&A on ratings using high pre-treatment rating variance (≥ 1) as a measure for fit mismatch.	42
2.8	The impact of Q&A on ratings using both low pre-treatment ratings (≤ 4) and high pre-treatment rating variance (≥ 1) as measures for fit mismatch.	43
2.9	The impact of Q&A on ratings using the pre-treatment fraction of reviews mentioning fit issues as a measure for fit mismatch.	44
2.10	Mechanism: products with a high fraction of pre-treatment fit related negative reviews experience a decline in such reviews following Q&A.	45
2.11	Review volume around the first answer.	45
2.12	Pageviews around the first answer.	46

2.13	The impact of Q&A on ratings controlling for price, discounts, and product description lengths.	47
2.14	Cosine similarity around the first answer.	48
3.1	Summary statistics of our main outcomes of interest.	83
3.2	Cramer’s V computed for covariates that had a statistically significant different in balance checks.	83
3.3	The effect of raising floor prices on user-levels outcomes: not including pre-treatment covariates.	84
3.4	The effect of raising floor prices on user-levels outcomes: including pre-treatment covariates.	85
3.5	The average user-level difference in dated and floor prices (Wedge) as a function of treatment status.	86
3.6	The impact of raising floor prices on users who have been exposed to dated prices.	87
3.7	The prices seen by users when they enter dates as a function of treatment status.	88
3.8	The impact of the wedge on outcomes on interest.	89
3.9	The impact of the wedge on outcomes of interest, instrumented by treatment status.	90
A.1	LDA topics extracted from the Q&A corpus.	105
A.2	LDA topics extracted from the review corpus.	106
A.3	Review volume following the first answer.	106
A.4	Pageview volume following the first answer.	107
A.5	Examples of reviews indicating fit and quality issues.	110
A.6	LDA topics extracted from the Q&A corpus.	111
A.7	LDA topics extracted from the review corpus.	111

A.8	Flexible definition of holdout: The impact of Q&A on ratings using low pre-treatment ratings (≤ 4) as a proxy for fit uncertainty.	112
A.9	Flexible definition of holdout: The impact of Q&A on ratings using high pre-treatment rating variance (≥ 1) as a proxy for fit uncertainty.	113
A.10	Flexible definition of holdout: The impact of Q&A on ratings using the pre-treatment fraction of review mentioning fit issues as a proxy for fit uncertainty.	114
A.11	LDA topics extracted from the pre-Q&A reviews.	115
B.1	A larger From price, as well as a smaller relative difference between the displayed From price and the actual price paid, leads to greater booking probability. The unit of analysis is a listing-scrape-calendar day. Column (1) uses Price and Price _{From} as independent variables; column (2) looks at their difference relative to Price and column (3) looks at their absolute difference	135
B.2	A larger From price, as well as a smaller relative difference between the displayed From price and the actual price paid, leads to higher occupancy rate (calculated as the fraction of total available calendar days that are booked).	137
B.3	Occupancy rate for “treated” listings with a positive change in min price (≥ 5) vs those with smaller or no change.	139

List of Figures

2.1	Descriptive features of UGC: We find that the accumulation of questions has risen steadily over time (in parallel with reviews), questions have been attracting answers faster over time, and most questions have a single answer.	49
2.2	Percentage of negative reviews (≤ 3 stars) that are due to fit-related issues by product category. (Limited to categories with at least 100 products and at least 100 reviews.)	50
2.3	The evolution of review volume for products with low ratings (top) and high variance (bottom), pre and post treatment (indicated by 0), measured 30 days around the first answer. The points plot the β_k coefficient estimates from Equation 2.12, and the bars indicate the 95% confidence interval. We see that there are no significant irregularities around the first answer time, thus mitigating the threat of omitted variables such as promotions.	51
2.4	The evolution of pageviews for products with low ratings (top) and high variance (bottom), pre and post treatment (indicated by 0), measured 30 days around the first answer. The points plot the β_k coefficient estimates from Equation 2.12, and the bars indicate the 95% confidence interval. We see that there are no significant irregularities around the first answer time, thus mitigating the threat of omitted variables such as promotions.	52

2.5	Plot of LDA topic weights for 20 topics, computed based on reviews that came from 0 up to 200 days prior to Q&A. We see no evidence of a change in review composition prior to Q&A arrival.	53
3.1	Example of advertised floor prices appearing in a Google search for ‘Hotels in Boston.’	91
3.2	A stylised model to depict the possible impacts of raising floor prices.	92
3.3	Example of experimental manipulation on Holidu.com.	93
3.4	Distribution of pre-treatment covariates across treatment conditions: no significant differences.	94
3.5	Distribution of pre-treatment covariates across treatment conditions: significant differences according to a Chi-Sq. test.	95
3.6	Distribution of pre-treatment covariates across treatment conditions: significant differences according to a Chi-Sq. test.	96
3.7	Comparison of point estimates and standard errors across OLS and doubly robust estimates. We find very consistent results.	97
3.8	Treatment effect heterogeneity across price levels.	98
3.9	Treatment effect heterogeneity across acquisition channel.	99
4.1	The ‘cost’ of leaving a review formulated as a threshold ρ . Our generative model simulates several histograms with different values of ρ_+ and ρ_- - the values here indicate the mean posterior ρ values computed across 939 products on Amazon.com.	102
4.2	Proportion across documents and topic-specific importance of keywords for the top 4 topics obtained by training an LDA topic model on question text across 63,000 books on Goodreads.com.	103
4.3	The proportion of books within the top 30 genres that receive at least one question. Each genre accounts for at least 1% of books in the sample.	104

A·1	Q&A technology on different platforms.	108
A·2	Screenshot of survey shown to workers on MTurk.	109
B·1	Region Country	128
B·2	Device	129
B·3	Host	130
B·4	Browser	131
B·5	Effect of the wedge difference in Prices and floor prices on booking probability. First, differences are binned in groups of 10. Then, for each bin, the average booking probability is computed. We see that the probability decreases as the difference increases. The bars indicate 95% confidence intervals, and the fitted line and shaded region plots out smoothed conditional means.	134

List of Abbreviations

ACM	Association for Computing Machinery
ARP	Advertised Reference Prices
IV	Instrumental Variables
LDA	Latent Dirichlet Allocation
ML	Machine Learning
NBER	National Bureau of Economic Research
Q&A	Questions and Answers
SVM	Support Vector Machine
UGC	User-Generated Content

Chapter 1

Introduction

The widespread expansion of online markets in the past decade poses several questions for platforms, firms and customers alike. Most of these questions revolve around the resolution of information asymmetries which are exacerbated in an online setting. In other words, to facilitate transactions, online markets need to make design choices that provide accurate information and build trust between buyers and sellers. My dissertation research is situated in this space and looks at a specific aspect of the information provision problem: given the increasing ease with which product pages can be customised to include a vast variety of content, how do these pieces of information interact? Further, what are the specific channels through which consumer decision-making is influenced by this? To study the above, I choose two important informational cues that most e-commerce platforms have adopted: (1) user generated content (UGC) and (2) advertised reference prices (ARP). Particularly, I try to understand how consumers respond to these cues, and how this might eventually affect the platform.

In chapter 2, I study how reputation systems (namely, user-generated reviews and ratings) on an e-commerce platform might be affected by a relatively new tool: Q&A technology. Q&As are also user generated – they allow customers to ask product-specific questions before purchase and receive answers either from other customers or the platform itself. I obtain data from a major UK-based e-commerce platform, and exploit variation in the timing of Q&As posted for various products to estimate their

impact on subsequent ratings and review text. I find that Q&As lead to higher product ratings by solving an important information problem relating to how consumers match with products in an online setting. Products often receive low ratings not because of vertical quality concerns but rather because of horizontal fit mismatches. An analysis of review and Q&A text reveals complementary roles for these two corpora: while reviews provide relevant information regarding the quality of a product before purchase, Q&As are much more effective at addressing idiosyncratic concerns about product fit that are specific to consumers. By alleviating such fit concerns, the addition of Q&As leads to better matches between customers and products, higher customer satisfaction, and result in higher ratings. Moreover, we build a text classifier to distinguish fit vs quality related negative reviews and find that the rating increase is driven by a reduction in reviews that highlight fit concerns.

In chapter 3, I aim to measure how consumers react when they are exposed to advertised reference prices that are not directly related to the price they need to pay. These prices are used widely across vacation rental platforms (which is the setting I study), and displayed when consumers conduct a search without entering dates. The particular ARP variant I consider is the ‘Starting from’ (floor) price, which has not been empirically investigated in the literature so far. In partnership with a large travel booking website (Holidu.com), I conduct two field experiments that exogenously vary the displayed floor price across users. Doing so, I find that users exposed to higher floor prices are less likely to engage with the platform - they make fewer outbound clicks, conduct fewer searches and spend lesser time on the website. They also tend to have a lower propensity to book and spend less on the website, although these booking related measures are not statistically significant due to sparsity. These effects occur despite higher floor prices providing users with a more accurate price estimate on average. I draw from the price obfuscation and salience

literature to justify these findings, and demonstrate that price salience (in the form of more accurate or transparent prices in this context) can in fact adversely affect consumer outcomes. Platforms thus need to carefully evaluate price display related decisions to optimise engagement.

Finally, in chapter 4, I provide some examples of research problems and ongoing projects that directly emerge from my dissertation work.

The primary tools I employ in my research are causal inference coupled with applications of machine learning. The paradigm for my second chapter is a quasi-experiment, whereas for the third it is a large scale randomised field experiment. Taken together, this research aims to demonstrate the possible consequences of heuristic cues that consumers are exposed to when they are shopping online, and provides insight into how firms might use this knowledge for optimal information provision.

Chapter 2

Interacting User Generated Content Technologies: How Questions and Answers Affect Consumer Reviews

This chapter studies the question and answer (Q&A) technology of electronic commerce platforms, an increasingly common form of user-generated content that allows consumers to publicly ask product-specific questions and receive responses, either from the platform or from other customers. Using data from a major online retailer, we show that Q&As complement consumer ratings and reviews: unlike reviews, questions are primarily asked pre-purchase and focus on clarification of product attributes rather than discussion of quality; answers convey fit-specific information in a predominantly sentiment-free way. Based on these observations, we hypothesize that Q&As mitigate product fit uncertainty, leading to better matches between products and consumers, and therefore improved product ratings. Indeed, when products suffering from fit mismatch start receiving Q&As, their subsequent ratings improve by approximately 0.1 to 0.5 stars and the fraction of negative reviews that discuss fit-related issues declines. The extent of the rating increase due to Q&As is proportional to the probability that purchasers will experience fit mismatch without Q&A. These findings suggest that, by resolving product fit uncertainty in an e-commerce setting, the addition of Q&As can be a viable way for retailers to improve ratings of products that have incurred low ratings due to customer-product fit mismatch.

2.1 Introduction

Consumer reviews have been shown to influence purchase decisions in the context of both products and services, and are widely adopted by brands and retailers (Chevalier and Mayzlin, 2006; Zhu and Zhang, 2010; Luca, 2016). Recently, another form of user-generated content, questions and answers (Q&As), has been gaining traction with online retailers and review platforms. Q&A technology, which is typically implemented alongside reviews, enables consumers to ask specific questions about a product and receive answers from another consumer, the brand, or the platform itself. Q&A technology is now widely adopted by retailers and has been embraced by consumers.¹ Despite this increased usage, little is known about how this technology affects consumer decision making.

In this paper, our aim is to fill this gap and examine the impact of Q&As on consumer decision making. Using data on consumer reviews and Q&As from a major UK-based electronic commerce platform spanning a period of 5 years, we show that Q&A technology resolves an important information problem and ultimately leads to better purchase decisions. The information problem arises from two sources of uncertainty that consumers face when trying to evaluate a product: quality uncertainty and fit uncertainty. Quality is a product-level characteristic that consumers agree upon, whereas fit captures idiosyncratic preferences that are specific to individual consumers. Fit uncertainty is exacerbated in an online setting because consumers cannot interact physically with a product prior to purchase.

We hypothesize that products often receive low ratings not because of quality concerns but rather because of consumer-product mismatches owing to fit uncertainty, possibly resulting from inadequate or wrong information on a retailer’s website, highly individualized fit requirements on behalf of consumers, or intrinsic product complex-

¹Appendix A shows examples of Q&As from various platforms. Amazon.com displays Q&As prominently above consumer reviews on each product page.

ity.²

To alleviate fit mismatch, online retailers have traditionally relied on consumer reviews. We posit that Q&A technology can act as an effective complement to reviews, and can help resolve any residual fit uncertainty that reviews might fail to address. By comparing our Q&A and review corpora, we find substantive differences in both how consumers use these information sources and their contents. We find that, unlike reviews, Q&As primarily happen pre-purchase, focus on clarification of product attributes rather than discussion of quality, and convey fit-specific information in a relatively concise and sentiment-free way. By contrast, because review text does not have a predefined structure that requires authors to comment on both quality and fit issues, it can be difficult for individual consumers to deduce fine aspects of product fit from this corpus. In these cases, Q&A technology can be a complement to consumer reviews since it allows individual consumers to inquire or read about their specific sources of uncertainty and receive answers before purchase. By addressing specific concerns about product features that may not come up in reviews³, Q&As can mitigate fit uncertainty before purchases happen. Thus, our main hypothesis is that Q&A technology can help resolve fit uncertainty where it exists, leading to greater consumer satisfaction post-purchase, which in turn results in higher product evaluations in the form of increased ratings.

A challenge that we face in testing our main hypothesis is identifying products that are more likely to suffer from fit mismatch. Because we do not directly observe whether a negative review is posted due to quality or fit related issues, we construct three different proxies for fit mismatch. Our first measure, motivated by the observation that mismatch causes low ratings, is the average rating of each product prior

²Please refer to Appendix A for some illustrative examples of reviews arising from poor quality vs poor fit.

³For instance, “My studio flat door is 27 inches wide, would it come through the door?” or “Does this work with Nikon L820 Bridge Camera?”

to the arrival of its first Q&A. While this measure is straightforward to compute, it is imperfect: some products might have low ratings due to quality issues rather than fit mismatch. Our second measure addresses this concern by taking into account the variance in ratings. High variance signals more heterogeneity in consumer preferences for certain attributes of the product, and therefore a higher likelihood of fit mismatch. Finally, our third measure takes into account the text content of negative reviews. We begin by asking human coders to read and categorize a sample of negative reviews into one of three categories: poor quality, poor fit, or other miscellaneous reasons (e.g., shipping concerns). We then use these human-labeled reviews to train a classifier that detects fit concerns. We apply this classifier to all negative reviews in our data to construct our third measure: the fraction of each product’s reviews that discuss fit issues.

Using data from a major UK retailer, we estimate the effect of Q&As on subsequent product ratings by exploiting variation in the timing of Q&As posted for different products. We find that answering questions for products that suffer from fit mismatch increases their subsequent ratings by roughly 0.1 stars. Moreover, we find that the extent of this rating increase is proportional to the probability that purchasers experience fit mismatch for that product prior to Q&A. To provide evidence around our hypothesized mechanism—that Q&As lead to better matches between consumers and products—we estimate the impact of Q&As on the probability of products receiving negative reviews mentioning fit-related issues. We find that the fraction of negative reviews due to fit concerns declines following the arrival of Q&As, with the extent of this decline again being proportional to the probability of fit mismatch prior to Q&A.

To interpret these results causally, we need to assume that the timing of Q&As is not correlated with time-varying unobservables that can also affect product rat-

ings. This assumption could be violated in our setting. In particular, unobserved marketing-related activities such as product page improvements, discounts, and promotions could attract more consumers to specific products, leading to Q&As. To the extent that these marketing activities are well-targeted, they could also lead to higher ratings. We approach these threats to validity in several ways. First, we use an auxiliary dataset of browsing behavior to directly look for patterns suggestive of demand shocks. We find no changes in the volume of reviews or product page impressions around the time Q&As arrive. Next, we collect additional data from the Internet Archive, which allows us to look at historical snapshots of the product pages in our sample. Based on this data, we re-estimate our main specifications controlling for historical prices, promotions and product description length, and we find no change to our results. These robustness checks suggest that our results are not being driven by unobserved marketing-related activities. Finally, we check whether there is an influx of reviews whose contents address fit concerns coinciding with the arrival of Q&As, which would confound our attribution. We test for changes in the composition of review text around the time of the first answer, and find no evidence that review contents change around the arrival of Q&As.

Overall, our findings suggest that, by resolving fit uncertainty in an e-commerce setting, the implementation of Q&A technology can be a viable way for retailers to improve product ratings, particularly for products that have suffered low ratings due to consumer-product fit mismatch.

2.2 Related work

The impact of ratings and reviews on consumer behavior (most notably, purchase decisions) has been well-documented in the literature. For example, in an online experiment, it was shown that participants who consulted product recommendations

selected these products twice as often as those who did not (Senecal and Nantel, 2004). Online consumer ratings have also been found to significantly influence product sales in the market for books (Chevalier and Mayzlin, 2006). Similarly, in the domain of services, a one star increase in a restaurant’s Yelp rating led to 5-9% increase in revenues (Luca, 2016). Although in this paper we do not directly look at purchases, these studies show that an increase in average ratings is a positive and managerially relevant outcome, since it has been widely shown to correlate with downstream conversion.

Other studies have looked more deeply into the impact of different characteristics of reviews on sales. More helpful reviews and highlighted reviews have been found to have a stronger impact on sales (Dhanasobhon et al., 2007). Further, the impact of reviews on sales is stronger for less popular products and for customers who have greater Internet experience (Zhu and Zhang, 2010). The text content of reviews has also been established to be of importance above and beyond the corresponding numerical rating (Archak et al., 2011).

In contrast to reviews, the role of user-generated Q&As in influencing conversion or related buyer behavior in an e-commerce setting has not been looked at. Most of the work in the domain of Q&As has focused on question-answering communities, such as Quora and StackOverflow. Questions examined in this area include: how reputation relates to response volume, question difficulty and answer quality on Stack Overflow (Lappas et al., 2017), how to model the satisfaction of information seekers in Q&A communities (Agichtein et al., 2009), what makes a “good” question in a community setting (Ravi et al., 2014), and so on. In terms of the interplay between Q&A-type communities and purchase behavior, it has been shown that engagement in firm-operated online communities can lead customers to spend more on the firm (i.e, accrue more “social dollars”), with this effect being strongest for posters of community

content, and those with more social ties (Manchanda et al., 2015). In such a setting, the source of economic benefit is seen as social rather than informational. Our work highlights an alternative channel through which Q&A platforms might provide a benefit to firms if they are integrated within an e-commerce framework, namely by resolving fit mismatch and leading to higher consumer satisfaction. The most closely related work to our paper, that also examines the overlap between Q&As and reviews in an e-commerce setting, develops an algorithm to show how existing reviews can be mined to answer questions on Amazon.com (McAuley and Yang, 2016). However, the focus of this work is developing and comparing the algorithm to other existing text mining tools, and does not investigate any causal questions that combine reviews and Q&As.

Our “Conceptual framework” section highlights how consumers make use of reviews and Q&As when both are present simultaneously on the product page. Closely related to the constructs of horizontal and vertical differentiation (Tirole, 1988), we posit that consumers are subject to two distinct varieties of uncertainty in an online setting: product quality uncertainty and product fit uncertainty. Broadly construed, product quality uncertainty is the consumer’s difficulty in evaluating product quality and predicting how a product will perform in the future (Dimoka et al., 2012). Products may have inherent quality issues that are revealed only through prolonged product usage - hence, reviews can be a valuable avenue through which quality uncertainty is mitigated.

Product fit uncertainty, on the other hand, arises because buyers cannot easily assess whether the product’s characteristics match their requirements or tastes (Hong and Pavlou, 2014; Kwark et al., 2014). Fit uncertainty might thus lead to mismatched purchases, and thereby attract low ratings even if the inherent product quality is good. While different consumers may have the same level of quality uncertainty

with a certain amount of information, their level of product fit uncertainty may vary due to their particular needs and heterogeneous fit preferences. Hence, we posit that attribute-based Q&A content can alleviate fit uncertainty more directly and completely than review text alone.

Prior work has also explored various other avenues through which these uncertainties can potentially be addressed, but without reference to Q&A technology. For instance, using survey data from consumers, it was seen that participation in online product forums reduces product fit uncertainty whereas the use of online media on product pages reduces product quality uncertainty (Hong and Pavlou, 2014). Our results relate to this work in the sense that we can think of Q&As as being similar to product forums that reduce fit uncertainty (in both cases, customers can bring up or read about specific concerns they have about a product). Fit uncertainty in an e-commerce setting can also be reduced with virtual reality widgets. For instance, in the context of apparel, it has been shown that offering virtual fitting rooms increases conversion, basket sizes, average price of purchased products, and revisits to the site, while reducing fulfillment costs arising from returns and home try-on behavior (Gallino and Moreno, 2018).

In the general domain of product returns, the two types of uncertainty have also been argued for: it has been shown that product fit uncertainty is mitigated by offline inspection and visits to the store, whereas reviews can offer a strong quality signal, thereby mitigating quality uncertainty, both of which can reduce return rates (Sahoo et al., 2018). We posit that, apart from offline inspections and augmented reality apps (e.g. virtual trials), Q&As can be an effective tool with which retailers can reduce fit uncertainty.

In addition to using average ratings and review text to measure the probability of inherent fit mismatch for a product, we also make use of rating variance. It has

been shown that niche products which some consumers like but others dislike can give rise to high variance (Sun, 2012). We would thus expect Q&As to facilitate better informed purchases for such products.

Finally, our setting differs from one in which quality and fit are more intrinsically linked and hard to disentangle (e.g. for books or movies). For instance, it has been shown that reviews on Goodreads.com can influence the nature of product discovery and thus shape consumer choices, by allowing consumers to find lesser known products that match their taste, more so than simply identifying products of high quality. In such settings, the role of Q&As would be more nuanced, since there are fewer objective attributes, and open-ended reviews might be more helpful in terms of resolving fit uncertainty (Bondi, 2019).

2.3 Conceptual framework

As summarized above, uncertainty in the context of e-commerce can be thought to be the result of two information problems: quality uncertainty and fit uncertainty. The industrial organization literature (Tirole, 1988) defines quality as a product-level attribute that is commonly perceived by all consumers, whereas fit reflects aspects of utility that are specific to individual consumers and can be highly idiosyncratic. In modern electronic commerce platforms, a key mechanism for reducing quality uncertainty is product reviews contributed by past purchasers. Product reviews can also provide information about fit. Nevertheless, reviews are not as well-suited to reducing fit uncertainty because they typically mix discussion of several aspects of a consumer’s experience with a product (e.g. performance, features, reliability, durability) using language that tends to be subjective and sentiment-laden (Liu et al., 2019). Moreover, the number of product attributes that relate to fit can be large and vary across consumers. Individual consumers may care about different subsets

of such attributes or may value the same attributes differently. For example, in the context of a smartphone, suppose that a consumer cares a lot about compatibility of the phone with an obscure hands-free protocol of an older vehicle. If no previous buyer of that product was interested in that exact product attribute, it is unlikely that any related information would be present, either in the product description or in the available product reviews. Q&A technology would enable that consumer to proactively ask a question about that, rather obscure, product feature and thus resolve her idiosyncratic fit uncertainty prior to purchase. In cases such as the above, Q&As act as a complement to reviews by allowing consumers to decrease their fit uncertainty prior to purchase. This in turn leads to consumers purchasing products better suited to their needs, and thus to fewer post-purchase regrets among those who choose to purchase.

The following hypotheses summarize the key predictions of our theory:

1. Q&As act as complements rather than substitutes for reviews since they are better able to provide fit specific information in a sentiment-free way.
2. The addition of Q&A alongside reviews is expected to resolve pre-purchase fit uncertainty. Resultantly, consumers are better matched to the products they purchase, which is reflected in higher product ratings.
3. The effect of Q&As is expected to be stronger for products that have higher residual fit uncertainty.
4. The addition of Q&As should result in a decline in negative reviews that arise due to fit concerns.

In Section A.2, we also present a mathematical model that captures how the presence of informative Q&A affects consumer decision making and product ratings in settings

with consumer fit uncertainty. Our model backs the above framework by showing that, if answers are reliably correct, Q&As increase ratings by eliminating fit mismatch. Further, the positive effect of Q&A on product ratings is proportional to the product-level probability that purchasers will experience fit mismatch without Q&A, which, in turn, is proportional to the amount of fit-related negative reviews without Q&A.

2.4 Data and descriptives

We obtain data from Bazaarvoice via the Wharton Consumer Analytics Initiative⁴. Bazaarvoice provides software that enables businesses to collect and display reviews and Q&As on their websites. Our data comes from a UK-based big-box retailer (similar to Amazon.com) that uses Bazaarvoice software. The data covers two product categories of consumer durables (Technology and Home & Garden) and includes all reviews and Q&As posted between 2009 and 2015. The two product categories are further subdivided into 755 subcategories such as Bedroom Furniture and Video Games. It is worth noting that for durables (unlike hotels/restaurants), it is less likely for firms to improve product quality over a short period of time, thus mitigating concerns of product quality increases in response to reviews.

Overall, our data contains 37,853 unique product identifiers.⁵ Out of these, 19,961 products do not have any user generated content (possibly because they were newly introduced products at the time of data collection) and thus cannot be considered for our analyses. Out of the remaining 17,892 products, 13,354 have at least one Q&A, 13,104 have at least one review, and 8,428 have both reviews and Q&As. Since

⁴<https://wcai.wharton.upenn.edu/>

⁵Some of these product identifiers refer to minor variations of the same underlying product, such as a white iPad and a black iPad, and share the same Q&As but have different reviews. We treat variations as independent products, because the ability of Q&A to alleviate fit uncertainty might differ across product variations, and aggregating them would lead us to underestimate the treatment effect of interest. Additionally, allowing for a product-level rather than a product-group level fixed effect in our estimation lends more flexibility to the model.

we want to study the impact of Q&As on product ratings, our analyses will focus on products that (1) have both reviews and Q&As and (2) have received at least one review before the first question was asked. This leaves us with 5,077 products, 345,168 reviews and 48,687 Q&A pairs. Table 2.1 presents summary statistics for all products in our estimation.

In addition to the above, we make use of click-stream data collected over a two month period in 2015 (February and March) to supplement our main analyses. This data consists of the browsing behavior of customers (i.e. which specific product pages were clicked on). Within this dataset, we look at products that received their first answer within the two month observation period.

We also collect data from the Internet Archive to conduct a series of robustness checks. These supplementary data sets are described in detail in the “Robustness checks” section.

2.4.1 Q&A, reviews, and fit uncertainty

The hypotheses we develop in this paper relate to the ability of Q&As to resolve fit uncertainty. In this section, we show that the Q&A corpus has a number of characteristics that make it particularly well-suited to conveying information about fit to consumers. We also compare Q&As to consumer reviews, and show that the two corpora differ in important ways that make Q&As better suited to resolving fit uncertainty.

We begin by examining the adoption of the Q&A technology, since the ability of Q&As to resolve fit uncertainty depends on the rate at which the feature is used by consumers. Figure 2.1 provides some descriptive patterns of Q&A dynamics. In Figure 2.1a and Figure 2.1b, we find evidence of increased usage of Q&As over time, mirroring the increasing adoption of reviews. In Figure 2.1c, we show that over time, questions have been getting answered faster: the average number of days it takes for

a posted question to be answered has gone down from 17 days in 2011 to 4 days in 2015, suggesting increasing engagement with the feature. Finally, in Figure 2.1d, we plot the distribution of answers per question. In our data, all questions are answered, and approximately 70% of questions have a single answer. We also find that close to 80% of answers to questions come from the platform itself, and not from other customers.⁶

In Table 2.2, we display the top product categories in terms of questions per product. We find that categories related to electronics and their accessories receive the most questions per product. Since these products are complex and typically associated with customer concerns about usability and compatibility, we would indeed expect a larger number of questions related to them.

Next, we report some descriptive evidence consistent with our hypothesis that Q&As mostly contain pre-purchase, fit-specific information, and do so in a less sentiment-laden fashion than reviews. First, to get at the pre-purchase nature of Q&As, we randomly sampled 2,400 questions. A set of 240 coders were then asked to classify a randomly chosen set of 10 questions each. We asked coders whether the question was most likely asked before or after purchase.⁷ We then computed (at the coder level) the fraction of responses that were in favor of questions being before purchase. We find that 83% of questions are posted pre-purchase. This is in stark contrast with reviews, which occur almost always (and for some platforms, exclusively) post-purchase.

To better understand content differences between reviews and Q&As, we perform a comparative sentiment analysis of the two corpora. We begin with a parts-of-speech classification of all reviews and Q&As. We find that reviews have a significantly higher

⁶This may not be the norm across different e-commerce platforms. Anecdotally, Amazon.com seems to attract more customer answers. Future work could examine potential differences in the effects of Q&As in environments dominated by customer vs. platform answers.

⁷For example, a pre-purchase question would be: “Is this keyboard compatible with MAC OS X Yosemite?”, whereas a post-purchase question would be: “My keyboard came with no instructions and the piece that raises the base already attached. How do I take it off?”

proportion of adjectives and adverbs (20%) compared to Q&As (9%). Adjectives and adverbs are known to be important components of sentiment analysis (Benamara et al., 2007). We also perform a sentiment analysis task on Mechanical Turk by asking coders to rate the sentiment content of 2,000 Q&A pairs (each pair is rated by two independent coders, with a third coder being assigned to break any ties), and find that 90% of them are rated as “neutral”. This leads us to believe that, while reviews are a more holistic expression of preferences, Q&As embody fit-specific information in a relatively sentiment-free way.

Finally, we examine the text of Q&As and reviews to gather additional evidence that Q&As are predominantly used for alleviating specific concerns related to product fit. To do this, we use a Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003).⁸ We train our LDA model on the entire body of reviews and Q&As and obtain 20 topics in each case. Consistent with our sentiment analysis above, the topics obtained for Q&As contain more references to fit-related attributes (such as dimensions, compatibility) than the topics obtained for reviews, which mostly contain information about product quality, or express sentiment in general. The top three topics (and associated highest-probability words) obtained in both cases are provided in Table 2.3.

2.5 Empirical Strategy

In this section, we begin by motivating our main estimating equation, and then discuss our identification strategy for measuring the causal impact of Q&As on consumer reviews. Our main hypothesis, articulated in the “Conceptual framework” section, is that Q&As can provide fit-specific information that allows better-matched

⁸We also build a Naive Bayes classifier that discriminates between Q&A and review text and find very similar qualitative conclusions: some of the top words that discriminate review content are “looks”, “value”, “money” and “great”, whereas that for Q&As are “helps”, “hope”, “confirm” and “using”. More details on this analysis are available upon request.

purchases, resulting in higher post-purchase utility and thus ratings. In our data, we observe individual reviews i left by purchasers of each product j , hence in what follows, we model purchasers rather than all consumers. This distinction is important since our theory does not predict that the same consumer i receives higher utility post-Q&A. Rather, the i th purchaser who selects into buying and rating product j in the presence of Q&A is happier than the counterfactual i th purchaser who buys and rates the product without Q&A and suffers mismatch.

We assume that products are both vertically and horizontally differentiated. Thus, each purchaser i derives post-purchase utility from two separate components of product j : vertical quality (η_j), which is product specific and common to all purchasers, and horizontal fit (μ_{ij}), which captures the degree to which j is a good match for purchaser i 's preference. Thus, ex-post utility takes the form:

$$U_{ij} = \eta_j + \mu_{ij} \tag{2.1}$$

To capture the potential impact of Q&As, we model the horizontal component of post-purchase utility as:

$$\mu_{ij} = \beta_j \cdot \text{POST}_{ij} + \epsilon_{ij}, \tag{2.2}$$

where POST_{ij} is an indicator set to one following each product's first answered question and zero otherwise, and ϵ_{ij} captures unobserved idiosyncratic factors that affect utility. To account for the fact that some products may suffer from fit mismatch more than others, we allow the effect of Q&As to be product-specific:

$$\beta_j = \beta_0 + \beta_1 \cdot m_j, \tag{2.3}$$

where m_j captures the degree of fit mismatch faced by purchasers of product j . Here, β_0 is an intercept term capturing the effect of Q&As on products facing no fit mismatch, and β_1 is a slope term capturing the effect of Q&As as the degree of

mismatch m_j increases.

Substituting Equation 2.3 into the utility function in Equation 2.1, we derive the following expression for the average utility obtained by purchasers of product j :

$$U_{ij} = \eta_j + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (m_j \cdot \text{POST}_{ij}) + \epsilon_{ij}. \quad (2.4)$$

If β_1 is positive, we can infer that on average, purchaser i is better matched with product j and therefore derives higher utility in the presence of Q&As. Further, the higher the degree of fit mismatch m_j , the greater the increase in utility. Under the assumption that ratings r_{ij} are an increasing function of utility, we modify Equation 2.4 above to arrive at the following estimating equation:

$$r_{ij} = \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (m_j \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \epsilon_{ij}. \quad (2.5)$$

Compared to the utility function above, this equation introduces additional controls, some of which depend on calendar time. We use the subscript $t(ij)$ to denote the year-month of review i for product j . Specifically, we model rating i left for product j as a function of product fixed effects η_j , time fixed effects $\delta_{t(ij)}$, time-varying observables X_{ij} , the POST_{ij} indicator, the fit mismatch term m_j , and unobservables ϵ_{ij} . In all specifications we estimate, we cluster standard errors at the product level (Donald and Lang, 2007).

The coefficient of interest, β_1 , has a causal interpretation under the assumption that the timing of each product’s first answer is as good as random. This assumption will be violated if increases in ratings and the propensity to answer questions are jointly driven by an unobserved process. In the “Robustness checks” section, we discuss and mitigate specific threats to this assumption.

2.5.1 Measuring fit mismatch

Our theoretical framework predicts that the effect of Q&A on product ratings is proportional to the degree of fit mismatch (m_j) that purchasers will experience without Q&A. However, the degree of fit mismatch inherent in any product's past purchases is not directly observable, since we do not know which negative reviews arise due to poor quality, fit mismatch, or other reasons. We approach this problem by constructing three proxies for the presence of fit mismatch prior to Q&A, which we describe below.

Low average ratings Products with a high pre-Q&A probability of fit mismatch will be purchased by many consumers for whom the product is a poor fit. In turn, these consumers will leave negative reviews for these products, leading to low average ratings. Hence, a simple heuristic for identifying products that suffer from bad fit is to focus on products with low average ratings prior to treatment (i.e. before the arrival of the first answer). We use an average rating of 4 out of 5 as the threshold that separates these products that may suffer from fit mismatch from those that don't. Later, we also show that our results are robust to different thresholds. The main concern with this measure is that fit mismatch is not the only source of negative reviews. Thus, by selecting products that have low average ratings prior to receiving Q&As, we may also incorrectly include products that do not suffer from fit mismatch (for example, low quality products). These false positives may attenuate the Average Treatment effect on the Treated (ATT) we estimate.

High rating variance Our second measure looks at products that have a high rating variance (Sun, 2012). We can think of such products as suffering from fit mismatch since they have attributes that are asymmetrically preferred by consumers (some like them and find them to be a good fit, others don't). We label products

whose rating variance is greater than 1 (the median) as suffering from mismatch. This measure also runs the risk of attenuation bias, albeit in a different sense: for products that have both high variance and high average ratings, bad fit might not be a dominant concern, and therefore Q&As might have less of an impact.

Thus, both the low ratings and high variance measures can misclassify products as suffering from poor fit when they do not. For example, when both the mean and variance of ratings are low, the most likely cause is poor quality rather than poor fit. Based on this observation, we also estimate specifications that combine these two proxies to identify products suffering from poor fit. Our expectation is that Q&As will be particularly helpful for low-rated high variance products.

Review text Our final measure looks at review text to identify products suffering from fit mismatch. To construct this measure, we build a text classifier that can distinguish negative reviews that arise due to poor fit. Using the classifier, we label each negative review as fit-related or not. Finally, we construct a continuous variable for fit mismatch for each product as the fraction of negative fit related reviews prior to each product’s first answer.

To build the classifier, we first construct a training set by asking two coders (on Amazon Mechanical Turk) to select most the likely cause of 3,300 randomly chosen negative (1-, 2-, and 3-star) reviews.⁹ We indicate three categories into which reviews are to be classified: poor fit, poor quality, or other miscellaneous reasons (such as issues with the store or shipping). Appendix A shows the survey seen by the coders. Any disagreements are resolved by a third coder. The coders classified 28% of negative reviews as having resulted primarily from poor fit and 67% primarily from poor quality. Since the third category accounted for a small fraction of the

⁹Refer to Appendix A for some illustrative examples of reviews arising from poor quality vs poor fit.

reviews (< 5%), relating mostly to in-store experiences and returns, we ignore it in our subsequent analysis.

We use these manually labelled reviews to train a C-Support Vector Machine (C-SVM) classifier (Cortes and Vapnik, 1995). To perform the classification task, we remove common stopwords, and then tokenize and stem the text of each negative review into a bag-of-words representation, thus obtaining word frequencies for each negative review. We use these word frequencies as predictors to train a classifier that predicts whether a negative review arises primarily from poor fit or poor quality.¹⁰

We train our classifier on an 80% random sample of our labelled data, holding out the remaining 20% to evaluate the classifier’s performance. The C-SVM classifier has one tunable parameter, C , which intuitively calibrates the trade off between classification accuracy and having a larger-margin separating hyperplane. We select a value for C using 5-fold grid search cross-validation. We evaluate the out-of-sample performance of our classifier using the commonly employed ROC-AUC (receiver operating characteristic area under the curve) metric. ROC-AUC ranges from 0 to 1 and it is a ranking metric. Intuitively, a ROC-AUC value of p implies a p probability of correctly predicting which of two reviews belonging to different classes (poor fit and poor quality) belongs to the poor fit class. Our classifier achieves a ROC-AUC of 0.82.¹¹

In addition to assessing the out-of-sample predictive power of our classifier, we also check whether our classifier makes qualitatively meaningful predictions about fit mismatch. We do so in three ways. First, in Table 2.4 we present the top-5 reviews with the highest predicted probability of belonging to each of the two classes (fit vs. quality issues). While reviews with a high predictive probability of being about

¹⁰We also considered using bigrams as predictors, but we did not see significant improvement in out-of-sample predictive power.

¹¹We also replicate our main analyses with a Naive Bayes classifier, which achieves an ROC-AUC of 0.79. These results are available upon request.

quality issues are explicit in mentioning poor product performance, reviews identified as having fit issues reflect customer-specific requirements that the product failed to fulfill, despite not inherently being of an inferior quality.

Second, we order all product categories in our dataset by the fraction of negative reviews that arise due to fit issues. We present these results in Figure 2.2. Intuitively, we would expect that categories for which a higher fraction of negative reviews are about fit would tend to be those for which it is harder for consumers to gauge whether the product is right for them. Indeed, we find that sofas (for which look and feel might be hard to gauge), electronic devices (which may involve compatibility issues) or accessories of some kind (which are meant to supplement a diverse set of existing items) tend to have more negative fit reviews. On the other hand, products where the customer’s domain knowledge dictates their purchase (such as DIY and power tools) have fewer fit concerns according to our classifier.

Third, we examine the top 20 words that are most predictive of fit vs. quality issues. To do so, we use layer-wise relevance propagation (LRP), a method originally developed to interpret the results of deep neural nets (Bach et al., 2015). LRP produces a score for each word and class (poor fit, and poor quality). High scores are assigned to words that are good at discriminating reviews belonging to each class. For linear SVM’s, the LRP score of each word-class combination is computed as the sum of the products of the word loading and the tf-idf score of the word in each of the reviews belonging to that class. We present the top-ranking words by LRP score for each of the the classes in Table 2.5. We find, as expected, that words which are a measure of objective quality (*work, poor, cheap, broke*) tend to be more predictive of negative quality reviews, whereas words that indicate more person-specific, idiosyncratic attributes (*look, need, design, however*) are predictive of negative fit reviews.

Overall, these results suggest that our text classifier can discriminate between reviews that bring up fit-related concerns and those that do not.

2.5.2 Mean reversion and measurement error

A final empirical challenge arises due to the fact that we construct proxies for fit mismatch as a function of past ratings, or, quantities correlated with past ratings (e.g., review text). This leads to two problems, which arise even if we assume treatment is strictly exogenous, i.e., $\mathbb{E}[\text{POST}_{ij} \cdot \epsilon_{ij}] = 0$. Here, we discuss these two problems under the assumption of treatment exogeneity; we discuss violations to treatment exogeneity separately in Robustness checks.

The first problem arises from applying a within transformation to Equation 2.5 to eliminate product fixed-effects η_j . The transformation mechanically introduces correlation between the demeaned residual and the demeaned version of $m_j \cdot \text{POST}_{ij}$, violating strict exogeneity and biasing OLS estimates of β_1 . (To see this, note that demeaning $m_j \cdot \text{POST}_{ij}$ and ϵ_{ij} introduces the mean error term in both quantities.) Although this type of bias is more commonly seen in models that explicitly incorporate a lagged outcome as a control (Nickell, 1981), it also arises in our setting because m_j is a function of lagged outcomes.

The second problem arises due to measurement error in the fit mismatch measure m_j . Recall that we do not observe m_j directly, instead relying on noisy measures $\tilde{m}_j = m_j + v_j$ (where \tilde{m}_j in our case could be mean ratings or rating variance), and v_j reflects unobserved factors uncorrelated with m_j that enter these proxies. For instance, some products may randomly experience transient shipping delays—a random shock to v_j —prior to their first Q&A leading to excess negative ratings. This could decrease the products’ mean ratings and increase rating variance, which we use as proxies for fit mismatch, for reasons unrelated to fit mismatch. Subsequent ratings for these products will likely revert back to their mean levels (e.g., once shipping delays are

resolved) regardless of any direct Q&A effect.

Rewriting our main estimating equation to reflect the use of proxies for fit uncertainty \tilde{m}_j , we have:

$$\begin{aligned} r_{ij} &= \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (m_j \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \epsilon_{ij} \\ &= \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot ((\tilde{m}_j - v_j) \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \epsilon_{ij} \\ &= \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (\tilde{m}_j \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \tilde{\epsilon}_{ij}, \end{aligned} \quad (2.6)$$

where

$$\tilde{\epsilon}_{ij} = -\beta_1 \cdot (v_j \cdot \text{POST}_{ij}) + \epsilon_{ij}. \quad (2.7)$$

Note that $\text{Cov}(\tilde{m}_j \cdot \text{POST}_{ij}, \tilde{\epsilon}_{ij}) \neq 0$, since both quantities depend on the unobservable v_j . This results in bias when Equation 2.6 is estimated by OLS.

We adopt a standard solution (Griliches and Hausman, 1986) to this classical measurement error problem, relying on a second noisy measure of our proxy, which we use as instrument. (Other papers that have used a similar empirical strategy in marketing include Narayanan and Nair (2013) - it has also seen widespread use in economics, e.g., Acemoglu and Finkelstein (2008), and Gupta (2021).) Specifically, we divide the pre-Q&A period for each product into two smaller samples: a hold-out period, which includes all reviews up to 200 days prior to the first answer, and a shorter pre-treatment period that includes all reviews starting at 200 days prior to the first answer and ending at the time of the first answer.¹² We then use these two samples to construct the two proxies $\tilde{m}_j^{\text{hold-out}}$ and \tilde{m}_j^{pre} , where the former quantity can be thought of as a lag of the latter. Finally, we use $\tilde{m}_j^{\text{hold-out}}$ to construct instruments for the endogenous variable $\tilde{m}_j^{\text{pre}} \cdot \text{POST}_{ij}$. The holdout sample is subsequently

¹²In a robustness check, we change the definition of our hold-out sample to make it more flexible: out of all pre-Q&A observations, we select the most recent 50% to form the pre-treatment sample, and the rest to form the hold-out sample. We report these estimates, which are similar to our main results in Appendix A

excluded from estimation. Our main estimating equation and the corresponding first stage are given by:

$$r_{ij} = \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (\widehat{\tilde{m}_j^{\text{pre}}} \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \tilde{\epsilon}_{ij}, \quad (2.8)$$

$$(\tilde{m}_j^{\text{pre}} \cdot \text{POST}_{ij}) = \tilde{\eta}_j + \tilde{\delta}_{t(ij)} + \gamma_0 \cdot \text{POST}_{ij} + \gamma_1 \cdot (\tilde{m}_j^{\text{hold-out}} \cdot \text{POST}_{ij}) + X'_{ij} \cdot \gamma_2 + \tilde{u}_{ij}. \quad (2.9)$$

where $\tilde{\epsilon}_{ij} = -\beta_1 \cdot (v_j^{\text{pre}} \cdot \text{POST}_{ij}) + \epsilon_{ij}$ and $\tilde{u}_{ij} = -\gamma_1 \cdot (v_j^{\text{hold-out}} \cdot \text{POST}_{ij}) + u_{ij}$. To see why this strategy works, notice that:

$$\tilde{m}_j^{\text{hold-out}} = m_j + v_j^{\text{hold-out}}, \quad (2.10)$$

$$\tilde{m}_j^{\text{pre}} = m_j + v_j^{\text{pre}}. \quad (2.11)$$

The instrument $\tilde{m}_j^{\text{hold-out}} \cdot \text{POST}_{ij}$ is valid under two conditions. First, it has to be relevant and have a strong first stage, which we can verify. Second, it has to satisfy the exclusion restriction $\mathbb{E}[(\tilde{m}_j^{\text{hold-out}} \cdot \text{POST}_{ij})\tilde{\epsilon}_{ij}] = 0$. This condition will be met as long as the two measurement errors are not correlated, i.e., $\mathbb{E}[v_j^{\text{hold-out}} \cdot v_j^{\text{pre}}] = 0$ (recall that we are assuming ϵ_{ij} is otherwise exogenous).

Intuitively, and continuing our prior example, we are assuming that products that experienced random shipping delays (and consequently excess negative ratings) in the pre-treatment period were not more likely to also experience such shocks in the hold-out period. Thus, by instrumenting with hold-out ratings we use the signal embedded in hold-out measures of fit mismatch (m_j) to get rid of the noise in pre-treatment measures of fit mismatch (v_j^{pre}), the latter being the source of bias when we estimate Equation 2.6 by OLS.

While we cannot directly test the exclusion restriction, we check whether ratings, which we use to construct proxies for fit mismatch, are serially correlated conditional

on observables. To do so, we conduct an autocorrelation test proposed by Arellano and Bond (Arellano and Bond, 1991; Roodman, 2009), and find that autocorrelation in levels vanishes beyond the first lag. Specifically, a rating may be serially correlated with the rating directly preceding it, but this serial correlation decays fast and is not statistically significant for the second lag and beyond. This provides us with some confidence that functions of ratings that are far apart in time (such as $v_j^{\text{hold-out}}$ and v_j^{pre}) are not correlated.

Using BLUP to construct instruments As we discussed above, the proxies we use for fit mismatch are measured with error. Beyond causing issues with identification, measurement error means that the instrument $\tilde{m}_j^{\text{hold-out}} \cdot \text{POST}_{ij}$ may be a poor predictor of $\tilde{m}_j^{\text{pre}} \cdot \text{POST}_{ij}$ for products with few reviews in the hold-out period. Here we explain how we obtain more precise measurements of our fit mismatch measures.

To obtain a stronger instrument we use a shrinkage estimator for $\tilde{m}_j^{\text{hold-out}}$, where we shrink the estimates of $\tilde{m}_j^{\text{hold-out}}$ for products with few reviews towards the population mean (Robinson, 1991). Specifically, for each product we estimate the best linear unbiased predictor (BLUP) of its mean rating in the hold-out sample using a mixed effects model with a random intercept m_j for each product j , and a fixed intercept μ :

$$\tilde{m}_j^{\text{hold-out}} = \mu + m_j + e_j$$

We apply this shrinkage estimator to the average rating and fraction of fit-related negative reviews instruments. Estimating the above equation, we obtain a BLUP of the mean rating and the mean fraction of fit related negative reviews in the hold-out sample for each product, which we use to construct our final instruments.¹³

¹³Results remain qualitatively unchanged even without employing the shrinkage estimator. However, they are slightly attenuated due to higher measurement error in m_j for products with fewer reviews. These results are available upon request.

2.6 Results

Now, we turn to estimating the effect of Q&A arrival on ratings using each of the three fit mismatch measures that we described above.

2.6.1 Low pre-Q&A ratings

Our first measure is low pre-treatment average ratings. Hence, we estimate Equation 2.8 with \tilde{m}_j^{pre} being an indicator for products with low average pre-treatment ratings (≤ 4), and $\tilde{m}_j^{\text{hold-out}}$ being the average rating of the product in the hold-out period. In addition to product and time fixed effects, we also control for the rank of each review (as recommended for example by Godes and Silva (2012)).

We present our results in Table 2.6. Column 1, which presents our OLS estimates, serves as the baseline and includes all products and all reviews in the estimation sample. We find a positive and significant increase in ratings of 0.24 for products with a low pre-treatment mean. We also find a statistically significant decrease of 0.045 stars for products that have a high pre-treatment mean. As discussed in the “Mean reversion and measurement error” subsection, some products may have high or low pre-treatment ratings by pure chance rather than as a consequence of poor fit. We would expect the ratings of these product to mean revert regardless of Q&As, which would inflate our estimates. In column 2, we re-estimate the OLS model by excluding the hold-out sample. Now, we see that both effects decrease in magnitude, but we still observe a small dip for products without fit uncertainty. Next, we move on to the IV specification. Column 3 reports the first stage of Equation 2.8. We see that average ratings computed based on the hold-out sample using BLUP are strong predictors of the pre-treatment average rating. Column 4 reports our preferred IV estimate: we find a positive and significant increase of 0.12 for low-rated products, and see no corresponding decrease for high-rated products. Finally, in column 5, to

capture treatment heterogeneity, we estimate a less parsimonious but more flexible specification where we interact the POST_{ij} variable with a full set of dummies for m_j being in different unit intervals (hence we do not include the main effect for POST_{ij}). We instrument each of these variables with the corresponding lagged version from the hold-out sample. We find substantial heterogeneity: the higher the fit mismatch, the larger a product’s post-treatment increase in ratings.

To put the magnitude of our effect—approximately 0.1 stars on average—in context, we compare it against the standard deviation of average ratings, which, for products with at least 5 reviews, is also roughly equal to 0.1 stars. The size of the effect we estimate is comparable to that of similar interventions such as the adoption of management responses (Proserpio and Zervas, 2017).

2.6.2 High pre-Q&A variance

Our second measure of fit mismatch is high rating variance prior to treatment. We estimate Equation 2.8 with \tilde{m}_j^{pre} being an indicator for products with pre-treatment variance ≥ 1 (the median value) and $\tilde{m}_j^{\text{hold-out}}$ being the rating variance in the hold-out period. We report our results in Table 2.7 for the full sample OLS (column 1), OLS excluding hold-out data (column 2) and IV (columns 3 and 4). In all cases, we find a positive and significant increase in ratings for high variance products following treatment. Similar to the previous case, a small decrease is observed for products with low variance, but this effect is very close to zero once we instrument for mean reversion.

As described in the subsection “Measuring fit mismatch”, we also estimate specifications that interact the low ratings and high variance proxies to better identify the set of products whose ratings have suffered due to fit-related concerns. This specification contains the full interaction between the rating and variance proxies resulting in four groups of products based on their pre-treatment ratings: low rating/low

variance, high rating/low variance, low rating/high variance, and high rating/high variance. We instrument each of these dummies with its hold-out equivalent. Given our theory, we expect low rating/high variance products to be primarily impacted by Q&As. We present our results in Table 2.8. Among the four groups of products, we see a statistically significant increase in ratings following Q&A arrival only for the low rating/high variance group, as expected. It is worth noting that we do not see a significant effect for high rated/high variance products. As previously mentioned, fit mismatch might not be the dominant concern for this group of products, since they are already high-rated.

2.6.3 High proportion of pre-Q&A fit-related negative reviews

Our final measure uses review text to detect pre-Q&A concerns about fit that might exist for a product. As a measure of the probability of fit mismatch, we compute the fraction of all reviews that are negative and fit-related prior to the arrival of the first question, based on the classifier described in “Measuring fit mismatch”. We estimate Equation 2.8 with \tilde{m}_j^{pre} being the fraction of pre-treatment fit related negative reviews and $\tilde{m}_j^{\text{hold-out}}$ being the fraction of fit related negative reviews constructed in the hold-out sample estimated using a logistic BLUP. We present our results in Table 2.9. The effect sizes for each specification mirror those found previously, thus indicating that the extent of rating increase is proportional to the fraction of fit related negative reviews. To illustrate how the effect sizes can be interpreted, consider the estimate in column 4: for a product with 10% of pre-treatment reviews expressing fit concerns, we estimate a subsequent increase in ratings of $1.135 \times 0.1 = 0.11$ stars.

2.6.4 Mechanism: fit mismatch reduction

We now turn to examining a hypothesized mechanism for our effect, namely that Q&As lead to higher ratings by promoting better matches between consumers and

products. To do so, we estimate the same specification as Equation 2.5, but with the dependent variable being an indicator for fit-related negative reviews (based on the classifier described in “Measuring fit mismatch”). We code all non-fit-related negative reviews and all positive reviews (4 and 5 stars) as 0. To match the definition of our dependent variable we use the text-based measure for fit mismatch (we obtain similar results for our two other measures, low ratings and high variance).

As before, we present results for both OLS and IV specifications in Table 2.10, and include a control for review rank. We find a negative and significant effect for the impact of Q&As on the probability of receiving a negative fit-related review for each of our specifications, indicating that the fraction of fit-related reviews declines following the arrival of the first answer, in proportion to the degree of fit mismatch prior to Q&A. Focusing on our preferred IV specification in column 4, we can interpret our estimates as follows: if 10% of all pre-treatment reviews are due to fit uncertainty, the product would experience a subsequent decline in the probability of receiving a negative fit-related review of -0.19×0.1 , i.e. 1.9%. This effect is relatively small due to the fact that the probability of receiving a negative review is low to begin with: in our data, only 15% of all reviews are negative (1-, 2-, and 3-stars.) However, conditional on receiving a negative review post-Q&A, the probability that this review is fit-related decreases by $\frac{100}{15} \times 1.9 = 12.6\%$.

Here we have shown that Q&As lead to fewer negative reviews that contain fit-related concerns. Our hypothesized theory for this reduction is that Q&As change the mix of consumers who purchase a product, i.e. Q&As affect selection into purchasing a product by helping consumers discover whether a product is a good match for them. However, Q&As might also affect who decides to leave a review. For instance, some consumers may make mismatched purchases because they neglected to read Q&As addressing their fit concerns. These consumers may later avoid leaving negative

reviews if they realize that the poor purchase was their own mistake. However, we believe that this is unlikely to be the prevalent mechanism for two reasons. First, it assumes that consumers who did not read Q&As when they were researching a product, decided to read them prior to leaving a review. While this is possible, we think it is unlikely. Second, if Q&As deter people from leaving a review, we might expect to see a reduction in the volume of reviews post Q&A arrival. However, this is not what we find (see Table 2.11.)

Robustness checks

Our results above indicate that answering a question leads to an increase in subsequent ratings for products that have suffered the consequences of fit mismatch. Moreover, we show that this increase is driven by fewer fit-related negative reviews post-Q&A. The primary threat to these findings is an unobserved time-varying confounder that drives both the arrival of Q&As and a subsequent increase in ratings, at any time in the post period, for products that suffer from mismatch. In this section, we investigate three such plausible confounders.

The first confounder we consider is promotions and discounts. Both increased advertising and reduced prices can increase demand for a product, resulting in more questions being asked and more reviews being submitted. The ratings associated with these new reviews may be higher than the product's current average rating due to lower prices, or due to a well-targeted advertising campaign that drives purchases from consumers who are likely to enjoy the product.

Next, we consider improvements to the product page. In response to a question being asked, the platform may update a product's description, which could alleviate fit uncertainty and thus increase ratings. In this scenario, while Q&A spurs the improvement of the product page, it is not the direct cause of increased ratings.

Finally, we consider an influx of fit-related reviews just prior to treatment. Here, it would be these new reviews that help consumers discover products that are better suited to their needs rather than the Q&A.

We address these concerns through a series of robustness checks. First, we show that there are no changes in review volume or product pageviews around each product's first answer, which we would expect in the presence of increased advertising. Next, we collect additional data that allows us to construct a panel of product descriptions, prices, and whether a product was being discounted. We find that our results are robust to controlling for price, discounts, and description lengths. Using the same dataset, we also show that the content of product descriptions doesn't change significantly around the time the first Q&A arrives. Finally, we check whether the content of reviews changes around the first Q&A and find this not to be the case. We describe these robustness checks in detail below.

Changes in review volume or pageviews A product-specific marketing campaign could raise demand for the product, leading to more Q&As, and if the marketing campaign is well-targeted, higher ratings. To guard against this concern we look for direct evidence that a marketing campaign may have been taking place around the time of each product's first answer.¹⁴ We focus on two outcomes suggestive of increased marketing activity: the daily number of reviews, and the daily number of pageviews each product receives.

First, we examine whether review volume increases significantly following each product's first answer. To do so, we estimate Equation 2.8 using the daily count of reviews each product receives as the dependent variable. As before, we instrument for fit mismatch using holdout measures and include weekday fixed effects as additional

¹⁴We focus on a 180-day period centred around the first answer, but our results are robust to other windows of time.

controls. We present our results using each of the three fit mismatch measures in Table 2.11. We find no significant change in review volume around the first answer. This also hints at the fact that while Q&As improve match, they do not have a direct effect on demand (proxied by the number of reviews) in the short term.

One concern with the analysis above is low power. Because reviews are relatively infrequent, a change in review volume can be difficult to detect. To increase power, we use click-stream data made available to us for a two month period (February and March 2015), and repeat our analysis using daily pageviews — a more frequent event — as our outcome. We estimate our regression using 1,091 products that receive their first answers during that period. We present our results in Table 2.12. We see no significant change in pageviews post treatment.¹⁵

Finally, we graphically examine any changes in review volume or pageviews in the immediate neighborhood of the first answer. To do so, we estimate the following model for our two discrete measures of fit mismatch, low rating and high variance:

$$y_{jt} = \eta_j + \delta_t + \sum_{k=-30}^{30} \beta_k \times \mathbf{I}\{D_{jt} = k\}_j + \epsilon_{jt} \quad (2.12)$$

where y_{jt} is respectively the number of reviews or pageviews and $\mathbf{I}\{D_{jt} = k\}$ is an indicator for day $k \in -30, 30$ for each product j . In Figure 2.3 and Figure 2.4, we plot the β_k coefficients from the volume and pageviews regressions, and observe no significant irregularities in pageviews or review volume.

Controlling for price, promotions, and product description length Our dataset lacks information on prices, and product descriptions over time. To deal with this issue, we download historical prices and product descriptions from the Internet Archive (IA), which is a non-profit digital library that collects historical snapshots

¹⁵We further address the issue of low power by using OLS on the full sample (results reported in Table A.3 and Table A.4 of Appendix A), and still find null effects. Arguably, estimating a precise zero effect even when mean reversion is present is a more stringent test of our hypothesis.

of web pages. We are able to find historical snapshots for 5,020 out of the 5,077 products in our data. In total we collect 145,564 snapshots of product pages, with an average of 29 snapshots per product. From each snapshot, we extract (a) the displayed price, (b) whether this price is marked down (based on the presence of the word “was” in the price field) and (c) the product description. We then associate each review in our sample with its chronologically nearest snapshot. We re-estimate our main specifications with three additional time-varying controls: prices, whether there was a price promotion, and the length of the product description. We find that our estimates for the impact of Q&A, shown in Table 2.13, are robust to the inclusion of these controls.

Changes in the text of product descriptions One concern with the above estimates is that character counts are a crude summary of product descriptions. Product description might remain the same length even though their content changes. Thus, we also investigate substantive changes in the contents of product descriptions within a 180-day window centered on each product’s first answer. To do so, we turn each product description into a bag-of-words representation: a vector of word counts scaled by each word’s inverse document frequency (a measure known as “tf-idf” in the natural language processing literature (Ramos, 2003)). To quantify changes in product descriptions over time, we choose each product’s first description as a reference point and compute cosine similarities between the first description and all subsequent descriptions. Finally, we estimate Equation 2.8 using these cosine similarities as our dependent variable. If product descriptions change following Q&A arrival, we would expect their cosine similarity to the reference description to decline. We report the results in Table 2.14. Overall, we see no decline in the similarity of product descriptions following the each product’s first answer.

Changes in review text One might be concerned that fit-uncertainty is resolved by consumer reviews that arrive alongside Q&As. This concern is partly mitigated by the fact that we control for the stock of pre-Q&A reviews received by each product using product fixed effects. However, an increased flow of fit-related reviews that coincides with the arrival of Q&As would bias our estimates. In this situation, we would expect to see a change in the composition of review text in the period leading to each product’s first answer reflecting an increased focus of reviews on fit-related issues. To investigate changing trends in review text, we begin by fitting an LDA topic model with 20 topics on all reviews prior to Q&As (the topics are available in Appendix A)¹⁶ We then group reviews based on their arrival relative to each product’s first answer. Finally, we calculate the average proportion of each topic within each group of reviews. We present these results in Figure 2-5. We observe that topic proportions remain relatively flat over time, leading us to believe that the contents of reviews do not abruptly change just prior to Q&A arrival.

2.7 Conclusion

In this paper, we study how the Q&A technology of e-commerce platforms affects consumer choice. We start by providing an overview of ways in which Q&A and review corpora differ, highlighting the differential ability of Q&As to resolve fit uncertainty. Moreover, we show that negative reviews might often arise due to consumer-product fit mismatch. Our main finding is that answering consumer questions can lead to a subsequent increase in ratings, which is driven by a reduction in negative reviews that arise from fit mismatch.

Overall, our findings have direct managerial implications for platforms, retailers as well as consumers. In the face of an ever-growing demand for information from

¹⁶The choice of 20 topics is motivated by coherence score measures (Röder et al., 2015). We obtain qualitatively similar results with different topic numbers that yield similar coherence scores.

online shoppers, it is important to realize potential positive synergies that might exist between different UGC features. Understanding these synergies can be important in guiding the platform’s adoption decisions when designing different interactive elements to integrate into the product page. From the perspective of retailers, Q&As might be an effective communication tool that directly allows them to interact with consumers before purchases happen. In the case of management responses to reviews, which also offer an avenue of retailer-consumer communication, it is often not possible to remedy the “damage” done by negative reviews, since this communication is post-purchase. However, Q&As can serve as an effective reputation management tool from that perspective, since they can serve to provide direct information that aids better purchases and mitigates the risk of negative feedback. Finally, from the standpoint of consumers, Q&As can lead to better informed purchases and hence higher post-purchase satisfaction, which can be expected to result in fewer product returns and more platform loyalty downstream.

Our results also have implications for the design of reputation systems. As we show, negative reviews can arise not just due to poor quality, but due to idiosyncratic fit mismatch. Platforms could consider running an algorithm similar to the classifier we propose, which could disambiguate these two broad classes of reviews, and compute a separate fit-based and quality-based average rating. This could be a possible way to mitigate the usual problems associated with biased average ratings on e-commerce platforms by leading customers to make more informed purchases.

A key limitation of our results is the inability to directly look at purchase behavior, since we do not have access to enough purchase level data. Some research has started to examine the impact of Q&As on purchases, and found evidence of a positive impact (Khern-am nuai et al., 2020). Future research could investigate how Q&As affect different stages of the purchase funnel (from consideration to purchase) and

the implications this has for firms and consumers. Another limitation is not being able to adopt a cross-platform identification strategy, which would have been able to more robustly rule out unobservable time varying shocks. We were unable to find a comparable platform that sells the same products but does not have Q&A.

It is also important to note that Q&As are not costless — there exists the possibility of potential information overload as more features accumulate on an e-commerce platform (such as videos and photos posted by users). Further, in our dataset, most questions have a single answer, but this is starting to change: for certain platforms, almost all questions receive multiple answers. This might give rise to ambiguities and perhaps change the objective/direct nature of the Q&A technology. Given these developments, it would be worth exploring the limits of the effect we observe, and better understanding what constitutes too much information (Branco et al., 2015).

On the whole, UGC implementations in general, and Q&A technology in particular, pose interesting problems for e-commerce websites that are worthy of further exploration.

2.8 Tables

Table 2.1: Summary statistics.

	Products	Avg. rating	Reviews	Questions
Technology				
Products with both Q&A and reviews	1175	4.37 (0.55)	153.6 (138.5)	23.78 (34.84)
Home & Garden				
Products with both Q&A and reviews	3902	4.28 (0.436)	264.5 (242.9)	14.2 (16.3)

Table 2.2: Top-10 product categories ordered by the average number of questions received.

Category	Questions Per Product
Set top boxes	42.79
iPod	23.84
Telephones and accessories	23.20
Televisions and accessories	19.40
DVD players	14.67
Fitted kitchens	13.94
Large kitchen appliances	13.51
Sat-nav and in-car entertainment	12.39
Heating and cooling	11.46
Kitchen electricals	10.83

Table 2.3: Top-3 LDA topics for reviews and Q&As.

Reviews	Q&As
Quality (of vacuum) easy, great, good, cleaner, clean, product, vacuum	Dimensions height, width, depth, dimensions, length, size, measurements
Quality (of electronics) sound, good, great, clock, quality, set, radio	Guarantee/Warranty buy, bought, guarantee, product, warranty, year, item
Quality (of phone/camera) phone, easy, set, good, camera, features, box	Compatibility (computers) ipod, compatible, laptop, windows, work, download, touch

Table 2.4: Top-5 reviews with the highest predicted probabilities of quality and fit issues.

Quality	Fit
This fan was not worth the money. It is poor quality for value. wasnt very efficient and broke after a couple of weeks. poor quality,some screws missing.	The boxes are small but quite strong. Quite small for the price Good quality - but too small for my requirement would have preferred it bigger
Poor quality, flimsy, and parts missing. Returned!	I use as back up to ecomy 7 it is a bit small though should have got two or a bigger one
This clothes dryer fell apart before I had even erected it. Very flimsy and poor quality. Took back next day!	Bit small for what I needed if for
Poor quality, ended up in bin.	Good Figures too small

Table 2.5: Top-20 words most predictive of quality/fit issues, estimated using layerwise relevance propagation on the output of an SVM classifier.

Quality	Fit
work	small
quality	look
poor	need
very	was
cheap	size
flimsy	does
return	fit
buy	just
good	design
money	colour
broke	shelf
make	use
any	however
week	bigger
day	did
screw	suitable
miss	picture
try	difficult
got	comfort
open	differ

Table 2.6: The impact of Q&A on ratings using low pre-treatment ratings (≤ 4) as a measure for fit mismatch.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1 st stage)	IV (2 nd stage)	IV (2 nd stage, bins)
POST \times Low Rating	0.243*** (0.022)	0.206*** (0.031)		0.122*** (0.041)	
POST \times Hold-out Rating			-0.901*** (0.029)		
POST	-0.045*** (0.009)	-0.026** (0.012)	4.049*** (0.130)	-0.011 (0.012)	
POST \times Rating $\in [2, 3]$					0.502*** (0.101)
POST \times Rating $\in (3, 4]$					0.100*** (0.037)
POST \times Rating $\in (4, 5]$					-0.012 (0.012)
Review Rank	0.0002*** (0.0001)	0.0001* (0.0001)	-0.0001* (0.00004)	0.0001* (0.0001)	0.0001* (0.0001)
Product FE	Yes	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes	Yes
F Statistic			312.91		
Observations	345,168	184,811	184,811	184,811	184,811

Note:

*p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 2.7: The impact of Q&A on ratings using high pre-treatment rating variance (≥ 1) as a measure for fit mismatch.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1 st stage)	IV (2 nd stage)
POST \times High Variance	0.159*** (0.015)	0.150*** (0.019)		0.085*** (0.030)
POST \times Hold-out Variance			0.533*** (0.016)	
POST	-0.061*** (0.010)	-0.053*** (0.012)	-0.064*** (0.019)	-0.026* (0.015)
Review Rank	0.0001*** (0.00004)	0.0001 (0.0001)	0.00002 (0.00004)	0.0001* (0.0001)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			380.08	
Observations	345,168	184,811	184,811	184,811

Note:

*p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 2.8: The impact of Q&A on ratings using both low pre-treatment ratings (≤ 4) and high pre-treatment rating variance (≥ 1) as measures for fit mismatch.

	IV (2 nd stage)
POST \times Low Rating \times Low Variance	-0.594 (0.429)
POST \times Low Rating \times High Variance	0.125*** (0.038)
POST \times High Rating \times Low Variance	-0.010 (0.014)
POST \times High Rating \times High Variance	-0.013 (0.028)
Review Rank	0.0001* (0.0001)
Product FE	Yes
Year-month FE	Yes
Observations	184,811

Note:

*p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 2.9: The impact of Q&A on ratings using the pre-treatment fraction of reviews mentioning fit issues as a measure for fit mismatch.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1 st stage)	IV (2 nd stage)
POST × Fit	1.932*** (0.150)	1.948*** (0.288)		1.135*** (0.356)
POST × Hold-out Fit			1.452*** (0.054)	
POST	-0.039*** (0.010)	-0.030** (0.012)	-0.007*** (0.001)	-0.013 (0.013)
Review Rank	0.0002*** (0.00005)	0.0001** (0.0001)	-0.00000** (0.00000)	0.0001** (0.0001)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			239.46	
Observations	345,168	184,811	184,811	184,811

Note:

*p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 2.10: Mechanism: products with a high fraction of pre-treatment fit related negative reviews experience a decline in such reviews following Q&A.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1 st stage)	IV (2 nd stage)
POST × Fit	-0.726*** (0.024)	-0.590*** (0.065)		-0.190*** (0.072)
POST × Hold-out Fit			1.452*** (0.054)	
POST	0.016*** (0.001)	0.013*** (0.001)	-0.007*** (0.001)	0.005*** (0.002)
Review Rank	-0.00002*** (0.00000)	-0.00001 (0.00000)	-0.00000** (0.00000)	-0.00000 (0.00000)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			239.46	
Observations	345,168	184,811	184,811	184,811

Note:

*p<0.1; **p<0.05; ***p<0.01

Standard errors clustered at the product level.

Table 2.11: Review volume around the first answer.

	IV (2 nd stage)	IV (2 nd stage)	IV (2 nd stage)
POST × Low Rating	-0.033 (0.071)		
POST × High Variance		0.020 (0.052)	
POST × Fit			-0.543 (0.732)
POST	0.094 (0.075)	0.080 (0.078)	0.100 (0.076)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes
Observations	25,257	25,257	25,257

Note:

*p<0.1; **p<0.05; ***p<0.01

Standard errors clustered at the product level.

Table 2.12: Pageviews around the first answer.

	IV (2 nd stage)	IV (2 nd stage)	IV (2 nd stage)
POST × Low Rating	3.708 (4.191)		
POST × High Variance		1.413 (1.133)	
POST × Fit			7.216 (9.394)
POST	-0.987 (1.630)	-0.632 (1.040)	-0.186 (0.819)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes
Observations	8,614	8,614	8,614

Note:

*p<0.1; **p<0.05; ***p<0.01

Standard errors clustered at the product level.

Table 2.13: The impact of Q&A on ratings controlling for price, discounts, and product description lengths.

	IV (2 nd stage)	IV (2 nd stage)	IV (2 nd stage)
POST × Low Rating	0.120*** (0.041)		
POST × High Variance		0.084*** (0.030)	
POST × Fit			1.111*** (0.352)
POST	-0.014 (0.012)	-0.028* (0.015)	-0.015 (0.013)
Review Rank	0.0001* (0.0001)	0.0001* (0.0001)	0.0001** (0.0001)
On Sale	-0.025** (0.011)	-0.025** (0.011)	-0.025** (0.011)
log(Price)	-0.141*** (0.027)	-0.140*** (0.027)	-0.142*** (0.027)
log(Desc. Length)	0.026 (0.016)	0.026 (0.017)	0.025 (0.016)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Observations	184,705	184,705	184,705

Note:

*p<0.1; **p<0.05; ***p<0.01

Standard errors clustered at the product level.

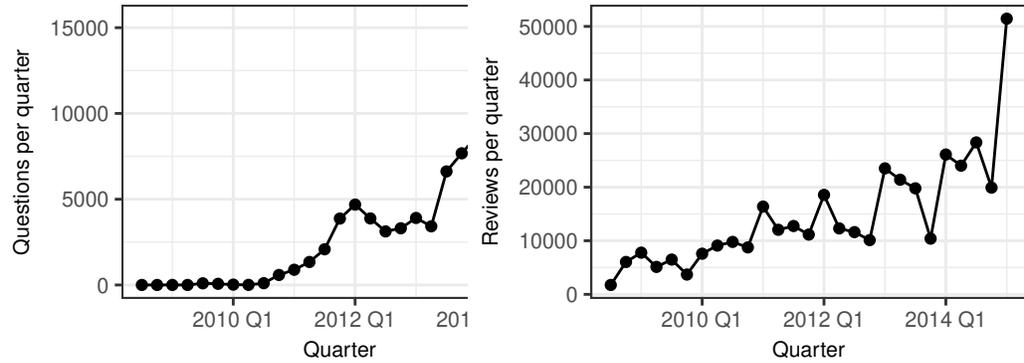
Table 2.14: Cosine similarity around the first answer.

	OLS (Rating proxy)	IV (2 nd stage)	OLS (Var. proxy)	IV (2 nd stage)	OLS (Fit proxy)	IV (2 nd stage)
POST × Low Rating	0.003 (0.005)	0.018* (0.009)				
POST × High Variance			0.008 (0.004)	0.013 (0.007)		
POST × Fit					-0.002 (0.033)	0.083 (0.067)
POST	-0.003 (0.003)	-0.006 (0.004)	-0.006 (0.003)	-0.007 (0.004)	-0.003 (0.002)	-0.004 (0.004)
Product FE	Yes	Yes	Yes			
Year-month FE	Yes	Yes	Yes			
Observations	20,432	14,741	20,432	14,741	20,432	14,741

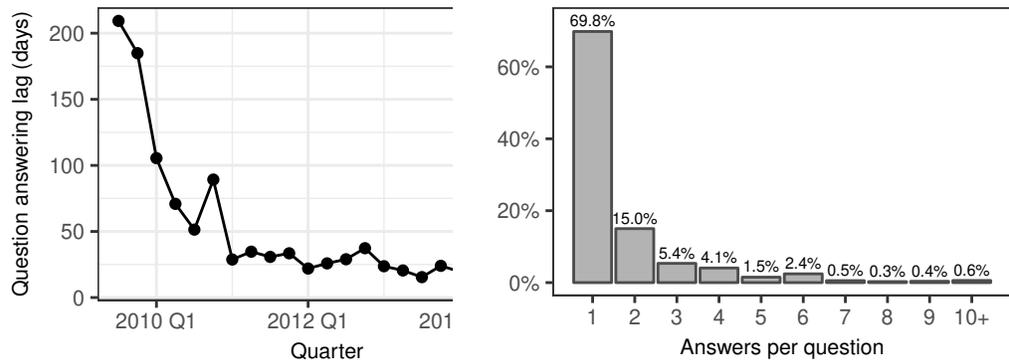
Note:

*p<0.05; **p<0.01; ***p<0.001
Standard errors clustered at the product level.

2.9 Figures



(a) Questions posted per quarter. (b) Reviews posted per quarter.



(c) Average lag between a question and its first answer by quarter. (d) Distribution of the number of answers per question.

Figure 2.1: Descriptive features of UGC: We find that the accumulation of questions has risen steadily over time (in parallel with reviews), questions have been attracting answers faster over time, and most questions have a single answer.

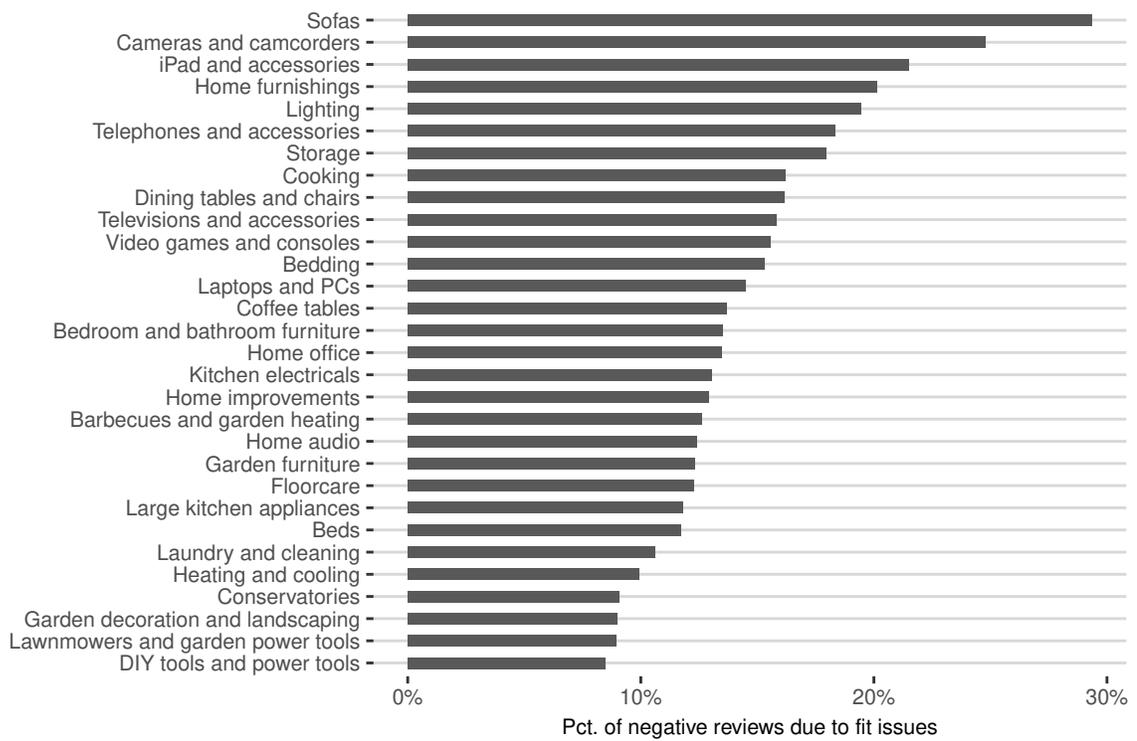


Figure 2.2: Percentage of negative reviews (≤ 3 stars) that are due to fit-related issues by product category. (Limited to categories with at least 100 products and at least 100 reviews.)

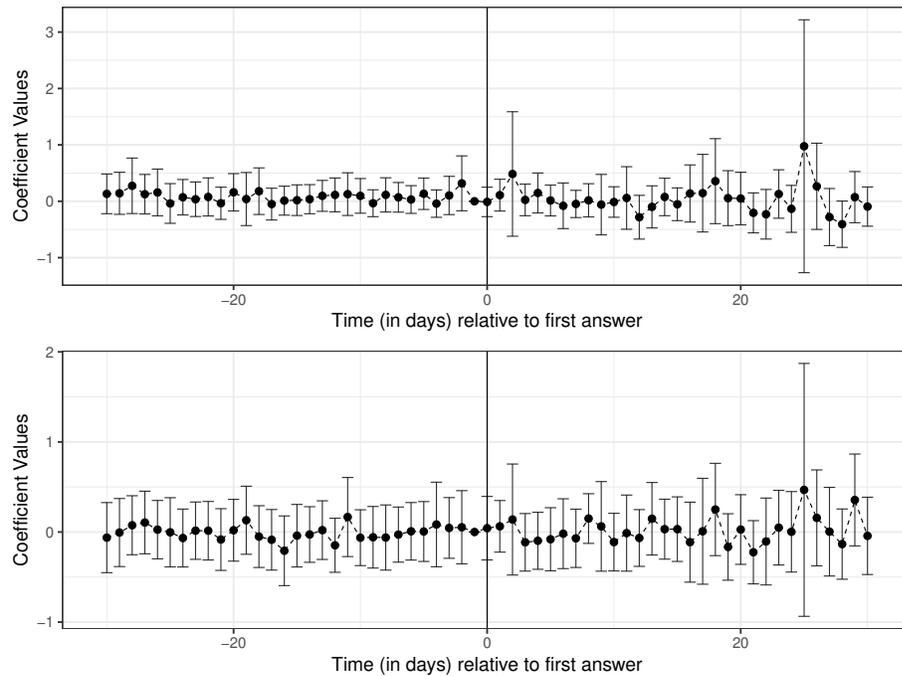


Figure 2-3: The evolution of review volume for products with low ratings (top) and high variance (bottom), pre and post treatment (indicated by 0), measured 30 days around the first answer. The points plot the β_k coefficient estimates from Equation 2.12, and the bars indicate the 95% confidence interval. We see that there are no significant irregularities around the first answer time, thus mitigating the threat of omitted variables such as promotions.

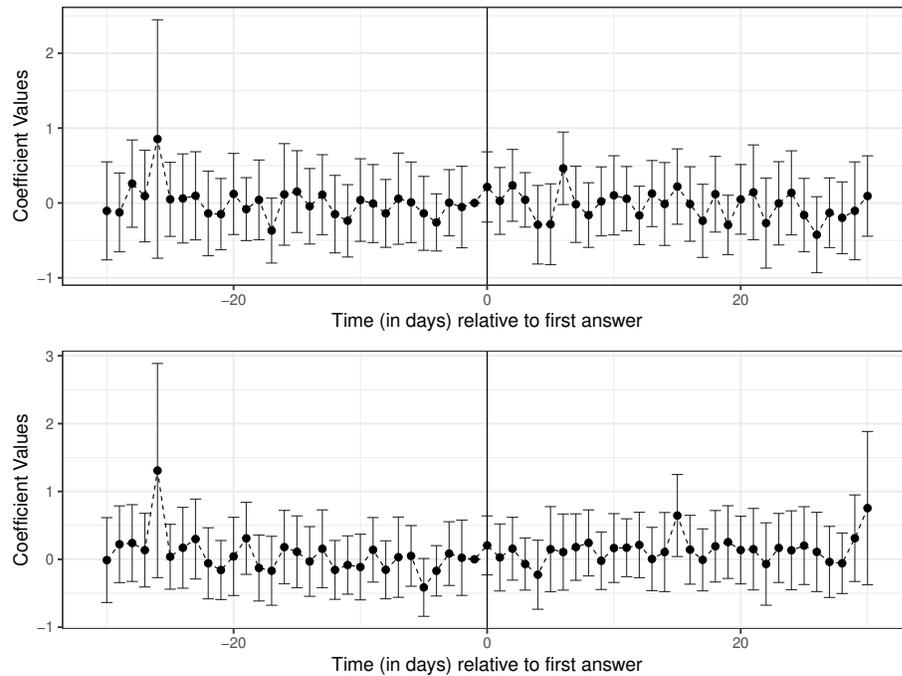


Figure 2-4: The evolution of pageviews for products with low ratings (top) and high variance (bottom), pre and post treatment (indicated by 0), measured 30 days around the first answer. The points plot the β_k coefficient estimates from Equation 2.12, and the bars indicate the 95% confidence interval. We see that there are no significant irregularities around the first answer time, thus mitigating the threat of omitted variables such as promotions.

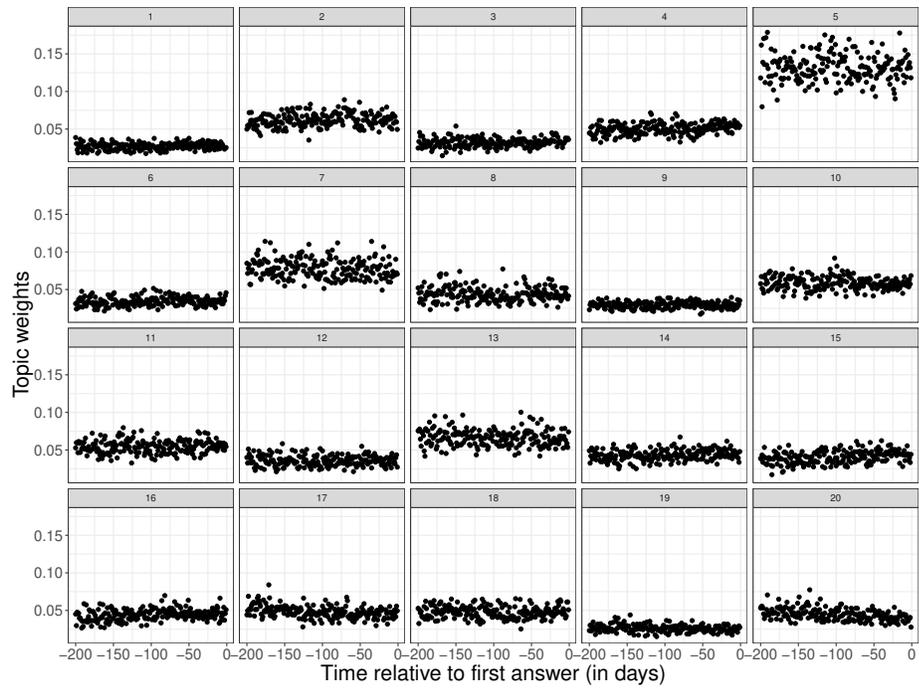


Figure 2-5: Plot of LDA topic weights for 20 topics, computed based on reviews that came from 0 up to 200 days prior to Q&A. We see no evidence of a change in review composition prior to Q&A arrival.

Chapter 3

Reference Price Effects In Vacation Rental Markets

Do consumers respond to prices advertised by firms, but irrelevant to the actual purchase price? A large literature on advertised reference prices (ARPs) has shown this to be true in various settings. However, ARPs are typically framed as discounts that serve to emphasize transaction utility. Our focus in this chapter is on “floor” ARPs that are particularly prevalent on vacation rental platforms. These prices appear in the form of a “Starting from”/“From” price when a user conducts searches without entering any dates. It is not clear how the accuracy of these price estimates relative to the true price might affect downstream outcomes. We thus conduct field experiments on a travel aggregator (Holidu.com) to investigate how consumers respond when “From” price estimates are raised. Our results indicate that higher floor prices actually lead to decreased engagement (as measured by outbound clicks, number of searches, and time spent on the website), and directionally negative effects on booking related outcomes. These effects occur despite higher floor prices providing users with an estimate closer to actual (i.e. dated) prices on average. We further demonstrate the possible moderating effects of acquisition channel and price levels. Overall, contrary to the sticker-shock theory wherein consumers are deterred if offered a low initial price estimate and a higher price further down the purchase funnel, we see that price obfuscation in the form of less accurate upfront prices can lead to more engagement. Platforms thus need to carefully evaluate the correct “balance” between optimising

customer engagement and providing accurate price estimates upfront.

3.1 Introduction

A reference price is defined as the standard against which the purchase price of a product is judged (Monroe, 1973). A sizeable literature in marketing, economics and psychology has established the various channels through which these prices can impact purchase likelihood, brand evaluations and other outcomes of interest. Reference prices that are formulated by consumers themselves (typically internal or external reference prices) usually rely on past purchase experiences, or prices of other comparable products. Thus, retailers cannot exert direct control over them. Nor can researchers directly observe such constructed reference prices, making it hard to conduct inference. On the other hand, advertised reference prices (ARPs) are supplied by sellers themselves, and can thus serve as a direct tool of comparative price advertising with which consumer price perceptions may be altered.

Typically, ARPs are presented in a "was-now" framing, or as a striked through price to emphasize apparent savings off a given list price. This is the framing that has received most attention in the literature. Several recent empirical studies have examined ARPs of this form, and found high ARPs to have a positive impact on purchase likelihood (Ngwe, 2018), and well as the likelihood of accepting the offered price without initiating negotiations (Jindal, 2018). The posited mechanisms for these effects is that high ARPs make the offer price more attractive, thus enhancing transaction utility. These effects are generally stronger for users with lesser category experience, who cannot precisely estimate the true market price of the item in question.

However, in the context of online marketplaces, platforms have a lot of flexibility in showcasing ARPs beyond just ceiling prices. With the vast majority of online prices being consumer-specific and customisable, retailers have to make decisions with regard to defaults, i.e, prices that are displayed before any specifics are entered. This decision is particularly relevant in the context of marketplaces or aggregators,

who are often themselves not price setters, but have complete control over the way price information is presented on the product page. These alternative ARPs have not received as much attention in the literature till now.

Our focus in this project will be on one such ARP variant that is increasingly being adopted by online marketplaces: namely, the “Starting from” price (henceforth referred to interchangeably as the floor price). For vacation rental platforms in particular, which is our setting, floor prices are ubiquitously showcased to consumers when they search for accommodations without entering any dates. They are also used heavily in the context of search engine optimisation by aggregators such as TripAdvisor and Booking.com - for example, a generic Google search for “hotels in Boston” contains floor price information in the returned results (see Figure 3.1). Despite the heavy adoption of these floor ARPs, there is no clear guidance on *how* these estimates should be constructed to optimise click-through, conversion, and other outcomes of interest. Should the window of prices over which the minimum is computed be selected in such a way that the lowest possible price is displayed, or is a higher (more realistic) floor price favourable?

Both floor and ceiling ARPs seek to make the offer price more attractive. However, for ceiling ARPs, the primary channel of influence is through a perceived sense of savings, i.e, emphasizing transaction utility, while the mode of operation for floor ARPs is unclear. Based on the literature, we hypothesize two primary mechanisms. The first possible mechanism draws from the “sticker shock” effect, namely that a large discrepancy between an upfront displayed price (which in our case is the floor price) and the true price leads to a negative surprise, thus adversely impacting firm outcomes (Winer, 1986). The second mechanism, however, would predict the opposite. Building on the theory of price obfuscation and salience (Ellison, 2005), this mechanism predicts that a lower floor price amounts to making prices less salient,

and will have a *positive* effect on subsequent outcomes.

Testing the impact of floor prices empirically in the context of online marketplaces is difficult with observational data, since any variation in displayed floor prices is likely to come from proprietary algorithmic changes that may be correlated with listing rankings. Thus, it is impossible to tease apart whether any effect on engagement is brought about by the algorithm as a whole, or by specific changes to the displayed floor price. Moreover, platforms are often unwilling or unable to exogenously manipulate displayed prices, which partially explains the limited number of field experiments in the literature that explicitly manipulate reference prices.

We are able to address these challenges by leveraging two unique price experiments with roughly 6 million users, conducted in partnership with Holidu.com (a Europe-based travel aggregator). Our treatment raises floor prices seen at the user level by 10% above the baseline. Further, the intervention occurs site-wide, thus alleviating concerns of non-experimental contamination. We find that in the treatment condition, engagement metrics such as outbound clicks, number of subsequent searches and time spent on the website are all affected negatively. Booking related outcomes are more noisy, but provide suggestive evidence: there is a directionally negative effect both on the propensity to book, as well as the total amount users spend on the website.

These results back recent empirical evidence from Blake et al. (2018), who demonstrate that displaying full prices upfront (relative to adding taxes and fees at the checkout page) decreases both the quantity and quality of ticket purchases made on Stubhub.com. While our data does not provide direct evidence for booking level outcomes due to sparsity, it complements these results by focusing on engagement metrics and demonstrating that consumer attrition occurs at every stage of the purchase funnel when higher floor prices are displayed. Thus, we find evidence in favor of price salience effects, and against sticker shock effects in our context.

Although higher “From” prices intuitively amount to providing more transparent price signals, it is an empirical question whether users are actually seeing more accurate prices in our experiment. To examine this, we narrow down on the set of users who conduct at least one search with a date, and are thus exposed to both dated (actual) and undated (floor) prices. We then compute the average user-level difference (henceforth referred to as the wedge) in these prices. We see that users in the treatment group see prices which are on average 10% closer to dated prices. Hence, higher floor prices do provide a more realistic estimate. However, even within this sample of consumers, outbound clicks continue to be affected negatively as a result of the treatment. We further examine how the absolute level of the wedge affects outcomes. Our experimental variation is no longer free of bias in this case, since dated prices are not randomly assigned across treatment and control groups. We thus rely on an instrumental variables strategy to estimate the Average Causal Response (ACR)(Angrist and Imbens, 1995) - we instrument the wedge with the binary treatment assignment variable to extract the part of the variation induced by random assignment. Doing so, we find that a higher wedge leads to positive outcomes - namely, a greater discrepancy between actual prices and floor prices actually leads to favourable outcomes, in line with our previous results.

Finally, we employ recent advances in causal machine learning to explore the space of conditional average treatment effects (CATEs), and to see whether we can detect meaningful heterogeneity in the treatment effects conditional of pre-treatment covariates. We use a doubly robust approach to compute individual level treatment effects, which are then projected on to the covariate space. This class of estimators produces a CATE which can tell us if there is heterogeneity in treatment effects for specific sub-groups formed based on observed pre-treatment covariates. Doing so, we broadly find negative effects across several different covariates. Interestingly

however, we find that users coming to the platform through Facebook search tend to spend more time on the website, and have directionally higher values of booking propensity and booking value. We also find that (1) the decline in outbound clicks in proportional to price level (defined by the average price of listings seen by a user), (2) booking metrics are negative and marginally significant for the lowest price level and (3) the reduction in engagement (time on site and searches) is affected more strongly at higher price levels.

Our paper contributes to the nascent but growing literature on online ARPs. More specifically, we are among the first to investigate the effects of displayed floor prices using a large-scale field experiment and real browsing behaviour. Our findings have implications for platforms and regulators who aim to provide consumers with more accurate price estimates upfront, and demonstrates that such a policy change might have unintended consequences.

3.2 Related work

Our paper relates broadly to literature on behavioral pricing and reference prices, as well as those on price obfuscation and fairness. In the following sections, we will provide a brief overview of existing work in each of these domains, and delineate our contribution relative to this literature.

3.2.1 Advertised Reference Prices

Transaction utility theory (Thaler, 1985) was among the first frameworks to show that buyers obtain some benefit simply from the perception that they paid less than their reference price. Subsequently, a large body of work has established the positive influence of firm-provided/advertised reference prices (ARPs) on perceived offer value (Urbany et al., 1988), purchase intent (Della Bitta et al., 1981; Bearden et al., 1984) and search intent (Della Bitta et al., 1981). The channel of influence is often through

internal reference prices (IRP) - researchers theorize that ARP is first assimilated into the IRP, which in turn influences purchase behavior or evaluations (Mazumdar et al., 2005). The ability of ARP to influence IRP is found to be affected by the plausibility of the ARP (e.g, Urbany et al. (1988)), as well as the difference between the ARP and the actual selling price (for instance, Kopalle and Lindsey-Mullikin (2003) find a U-shaped relationship between advertised reference prices and consumer price expectations). It is also influenced by semantic cues (e.g., was-now versus compare at) that retailers use to frame the sale (Lichtenstein et al., 1991). In addition to the above, plausibly high ARPs have also been shown to increase internal reference price standards, and thus enhance fairness perceptions of the offered price (Lichtenstein and Bearden, 1989; Xia et al., 2004). While most of these papers make use of lab experiments, Mayhew and Winer (1992) were among the first to use actual transactions data to study the role of ARP in affecting consumer choices.

There has also been a recent upsurge of papers utilising real transaction data and non-survey based experimental design. For instance, Huang (2018) utilizes lab experiments to show that high ARPs provide consumers transaction utility which increases their likelihood to purchase. Ngwe (2018) utilises transaction data on a clothing retailer and shows that “fake discounts” in the form of high list prices (with the actual price being marked down relative to that list price) have a strong influence on purchase outcomes, with a \$1 increase in the list price having the same positive effect on purchase likelihood on average as a \$0.77 decrease in the actual selling price. This effect is larger for fake list prices, but smaller in longer-lived stores and stores closer to regular retail channels. Jindal (2018) also uses purchase data and shows that high ARPs increase the negotiated price of big ticket items.

However, to the best of our knowledge, the bulk of work in this space has focused on ceiling ARPs that are framed as a promotion. There has been lesser focus on

floor ARPs and other unconventional price estimates that are prevalent in online marketplaces. We are also among the first to conduct large scale field experiments that systematically manipulate floor prices and examine how user interactions with the platform evolve as a result.

3.2.2 Price obfuscation and salience

The decision of displaying low floor prices can also be looked upon as a form of price obfuscation, given it can lead consumers to underestimate the eventual price they have to pay. Our work therefore contributes to the literature on digital obfuscation strategies and their possible effects. Broadly, obfuscation can be thought of as an action that raises search costs, which can lead to less consumer learning and higher profits (Ellison, 2005; Ellison and Ellison, 2018). Another way to think about obfuscation is in relation to sales of “add-ons” at high unadvertised prices, which can raise equilibrium profits in a competitive price discrimination model. Designing products to require add-ons can thereby be a profit-enhancing obfuscation strategy even when consumers correctly infer all prices. In fact, even when consumers are aware of add-on pricing policies, they prefer to give their business to firms who “shroud” prices because these sophisticated consumers end up with a subsidy from policies designed for myopic customers (Gabaix and Laibson, 2006).

In a different domain, Chetty et al. (2009) demonstrates how lesser obfuscation (in the form of tax salience) can have a negative impact of demand. Specifically, the paper shows how commodity taxes that are included in posted prices reduce demand significantly more than taxes that are not included in posted prices. The fact that individuals make such optimization errors even with relatively simple, linear commodity taxes suggests that more complex policies such as income taxes or transfers could generate very different behavioral responses from those predicted by standard models.

Several recent studies have further explored possible gains and related caveats from obfuscation strategies in digital settings. For instance, Allender et al. (2021) find in an online experiment that price obfuscation is highly effective at mitigating consumer peer-induced fairness concerns, in turn raising the average price charged. They further find that buyers are more likely to make a purchase when the prices are obfuscated even though they knew the seller had intentionally, and strategically, reduced price transparency. Sellers may use strategic obfuscation to avoid fairness-based barriers to individual pricing, without the need to negotiate prices. However, there is a trade-off between the peer induced and distributional fairness concerns: once the prices are obfuscated consumers shift their attention to evaluate the distributional inequity more scrupulously.

Mamadehussene (2020) demonstrates that there is tension between platforms and firms regarding how much price complexity is used: firms would like to use even more obfuscation than what the platform allows, so the platform must monitor firms' prices to make sure that they are not excessively complex.

Finally, Blake et al. (2018) find that price obfuscation on Stubhub (i.e, having back-end fees) leads to higher revenue relative to upfront prices. Detailed click-stream data shows that obfuscation makes price comparisons difficult and results in consumers spending more than they otherwise would. Consumers who are shown fees upfront drop off early in the purchase funnel, while those shown fees later are more likely to exit after the site displays total prices, consistent with consumer misinformation. However, salience persists beyond initial misinformation. Experienced users, who arguably should anticipate the fee, spend 15% more on StubHub when the fee is shrouded. This behavior suggests that experience with Back-end Fees does not give users an advantage in anticipating true final prices. In general however, the paper notes considerable heterogeneity in the pricing practices of platforms, which points to

the need for more empirical work that can tease apart the effects of price obfuscation in a digital setting.

Our results draw from the obfuscation and price salience literature and demonstrate an empirical application of the positive impact of obfuscation strategies on engagement metrics. By doing so, our paper contributes to studies of alternative methods of obfuscation, such as add-on pricing and partitioned pricing. Similar to Blake et al. (2018), exact (i.e. dated) prices in our setting constitute surcharges rather than add-ons because they are unavoidable. We might interpret the displayed per night price as a form of partitioned pricing (Morwitz et al., 1998) where the base (floor) price is essentially augmented by the nightly surcharge, depending on temporal demand and other factors. One interpretation of our findings is that price salience amplifies the effect of partitioned pricing.

3.2.3 Fairness perceptions

Our work also relates to the literature on ambiguous price claims and price fairness. Central to any model of price fairness is the notion that buyers, either explicitly or implicitly, have some sort of reference price they use to assess whether or not a price is fair.

Previous research on price fairness shows that a perceived price discrepancy might cause negative fairness perceptions for consumers (Xia et al., 2004). Relatedly, this can also be a form of deceptive advertising. Lab studies have demonstrated that while deceptive advertising can increase false brand attribute beliefs (Burke et al., 1988), it can engender distrust which negatively affects response to subsequent ads (Darke and Ritchie, 2007). Papers have also demonstrated negative effects of buyer antagonism. For instance, Anderson and Simester (2010) find in a field experiment that customers react by making fewer subsequent purchases if they buy a product and later observe the same retailer selling it for less.

In the context of hotel resort fees, Sullivan (2017) finds that separating mandatory resort fees from posted room rates without first disclosing the total price is likely to harm consumers by increasing the search costs and cognitive costs of finding and choosing hotel accommodations. The analysis finds this strategy is unlikely to result in benefits that offset the possible harm to consumers. Interestingly, however, our findings suggest that consumers are driven less by fairness perceptions than they are by the upfront price displayed on the website. Although higher floor prices are more accurate, and thus more fair, consumers engage with the platform less in our specific setting.

3.3 Conceptual framework

Our goal is to demonstrate the impact of high floor prices on engagement metrics such as outbound clicks, number of searches and time spent on the website, as well as downstream metrics like bookings and booking value. Based on the literature, we hypothesize that displaying higher floor prices can have one of two effects. The first possible mechanism draws from the “sticker shock” theory, namely that a large discrepancy between an upfront displayed price (which in our case is the “From” price) and the true price to be paid leads to a negative surprise, thus adversely affecting product level outcomes (Winer, 1986). A smaller sticker shock is also plausibly related to greater fairness perceptions among consumers (Xia et al., 2004) - high floor prices can lead consumers to anchor their expectations upwards, which leads to a smaller discrepancy with the actual price, thus stimulating them to trust the platform and engage more. If this is the dominant mechanism operating in our context, we would expect higher floor prices to lead to more favourable outcomes.

However, the second posited mechanism builds on the theory of price obfuscation and salience (Ellison, 2005; Chetty et al., 2009) and yields the opposite prediction.

Obfuscation has been theoretically and empirically shown to raise firm profits, while salience (e.g. posting tax-inclusive price tags rather than applying taxes at the register) reduces demand. These effects have been shown to persist even when consumers are aware of prices being obfuscated. This mechanism predicts that floor prices are an implicit tool for obfuscation. Under that assumption, a higher floor price amounts to increasing the salience of true prices, and will have a *negative* effect on subsequent outcomes.

To formalize our intuition, we build on a stylised model developed by Blake et al. (2018), and illustrated in Figure 3.2. Let J be the set of hotel listings faced by consumers under the default floor price regime. The consumer's optimisation problem is to make a quality-price tradeoff as they are forming their consideration set and deciding whether to engage with the platform. For illustrative purposes, we will focus on the consumer's decision to click. However, similar arguments follow for each of our examined outcomes. The optimisation problem can be formulated as:

$$v_i = \theta q_i^{true} - p_i^{true} \quad (3.1)$$

where i is the listing chosen to maximise v_i . q_i^{true} and p_i^{true} denote the true quality and price level that correspond with the choice of i . Following Blake et al. (2018), θ captures the trade-off between quality and price. Higher values of θ indicate a steeper indifference curve whereby consumers are willing to pay more for greater quality.

The outside option is denoted by 0, with $q_0 = p_0 = 0$. The straight line $v_0 = 0$ marks the consumer's indifference curve from not interacting with the site. Given J , the consumer chooses p_i^{true} and q_i^{true} to maximise utility $v' > 0$. For consumers with low enough values of θ (less steep indifference curves in Figure 3.2), their indifference curve $v_0 = 0$ lies fully below the set J , and they will not click out. It therefore follows that given a set of hotels J , there exists a threshold type $\theta' > 0$ such that a consumer

of type θ will click out if and only if $\theta > \theta'$.

When consumers conduct a search without dates, they do not see p_i^{true} but a p_i^{floor} , which is a noisy signal of p_i^{true} . Building on the reference price literature, we assume that the perception of p_i^{true} is influenced by p_i^{floor} seen by consumers. Let the perceived price be \tilde{p}_i^{true} .

The optimisation problem is now to choose i such that \tilde{v}_i is maximised:

$$\tilde{v}_i = \theta q_i^{true} - \tilde{p}_i^{true} \quad (3.2)$$

When p_i^{floor} is raised, the following scenarios may occur:

1. Price salience theory would predict that consumers shift their perception about \tilde{p}_i^{true} upwards. This effectively shifts the choice set J to the left (to the *perceived* choice set J'), whereby consumers perceive that they have to pay a greater price for the same q_i^{true} levels.
2. On the contrary, the sticker shock theory would suggest that consumers derive positive utility from accurate price estimates, and a large discrepancy creates disutility (salience models typically do not account for such effects). In this case, raising p_i^{floor} amounts to calibrating more accurate price expectations, as a result of which consumer trust in the platform, and quality assessment increases. Resultantly, the consumer perceives *greater* quality at the same p_i^{true} levels. Hence, the perceived J shifts right to J''.

Following the intuition of Blake et al. (2018), a leftward shift of the choice set implies a negative impact on choice, whereas a rightward shift implies a positive impact. This is because, the set of consumers with $\theta < \theta'$ will prefer not to click if they perceive the set of listings to be J. Some of these consumers, however, will select

a listing if they perceive the choice set to be J' . Conversely, some consumers who would have clicked under choice set J will choose not to do so under choice set J' .

Our empirical results find support for the price salience theory - in other words, raising floor prices leads consumers to perceive the “true” choice set as J' , and deters a subset of consumers from engaging with the platform, with indications of possible negative effects on bookings and booking value too.

3.4 Experimental details

3.4.1 Experiment design

Our aim is to examine the effects of raising floor ARPs on consumer level browsing and booking behavior. Consistent with this aim, our unit of randomisation is at the user (cookie) level. This ensures that a specific user-device combination gets assigned to the same treatment condition every time they conduct a search through the duration of our experiment, as long as they do not clear out their cookies. Floor prices are visible to any consumer who conducts at least one search without a date (this condition is satisfied by 97% of our sample). Each time a unique user visits the website and conducts a non-dated search, they are assigned with equal probability to either the baseline¹, or the higher floor price (treatment) condition. Each search also triggers an impression event, which gives us detailed information on the specific listings that the user is exposed to for a given search, and the associated prices they see. For our subsequent analysis, we will aggregate all search-level metrics at the user level.

What makes our experimental setting different from most extant literature on reference prices in general, and ARPs in particular, is that we can exogenously vary

¹Baseline floor prices are computed by the platform using cached information for each listing, e.g. previous prices, or price information supplied by the listing itself. For our experiment, the exact algorithm used to calculate the baseline does not matter, since we uniformly raise it by 10% in the treatment condition.

these prices, and thus measure their role in affecting outcomes of interest. An alternative research question could have been to examine the role of the “From” price widget as a whole. To do this, we would need to expose users to a version of the website with floor prices, and one without. This would tell us whether explicitly helping consumers to construct their price expectations, rather than relying on their autonomous expectations based on experience, competitors, etc serves the platform better. However, we have no way of knowing consumers’ autonomous expectations and how they are constructed. In the setting of a field experiment, there is no natural way to solicit these internal expectations either. Given these constraints, we feel that our design (and associated conclusions) are more managerially relevant.

3.4.2 Data

Our data comes from Holidu.com, a Europe-based travel aggregator, with close to 15 million listings across the world.² Our experiments are conducted in two waves, in 2019 (March 5th to April 23rd) and 2020 (January 29th to March 13th). Each time, the experimental manipulation is applied site-wide, thus mitigating concerns of non-experimental interference. Power calculations were conducted based on results from the first experiment, and the second experiment is pre-registered at asPredicted.org.³ There were 6,979,342 users in total who were exposed to floor prices, searching for accommodation across 35,950 cities in 89 countries. These users originated from 37 unique domains, which serve as a rough proxy for origin country locations. As mentioned before, the treatment status of a user (as identified by cookies) remains

²<https://www.holidu.com/>

³We originally intended to use data only from the follow-up experiment and not include the pilot results. Unfortunately, due to the global coronavirus pandemic, we had to halt our follow-up experiment before attaining an adequate sample size (roughly 7 million) to detect a 5% change in our booking metrics (propensity to book and booking value). Resultantly, we pooled together results for both experiments to give us a bit more precision, but still not enough to detect changes smaller than 5% for the booking metrics. However, despite not attaining statistical significance, the directional negative effects on bookings described in the Results section provide complementary evidence for the likely implications of raising floor prices.

unchanged during the duration of the experiment, so we can observe whether each user conducts multiple searches, and the specific impression events triggered during each search. A total of 31.9 million searches are conducted by these users during the experiments, which are aggregated at the user level for our analysis. 63.1% of searches lead to users browsing a single page of the returned results (i.e, one page of impression events) - 97.02% of all searches remain within the first 10 pages of impressions.

Users are randomised into one of two floor price conditions: baseline and baseline + 10% (example in Figure 3-3). In the 2019 version of the experiment, a third condition (baseline + 5%) was also used, but we do not use results from that group for consistency. Resultantly, we are left with a total of 6,385,717 users. Participants are split virtually equally into the two conditions, thus indicating that our randomisation was successful. We look at six main outcomes: outbound clicks, total number of dated searches, total number of overall searches, the likelihood of booking, booking value and total time spent on the website (measured as the timestamp difference between the first and last searches conducted by the user during the experiment). These variables are summarised in Table 3.1. We also collect data on user-level covariates, namely acquisition channel, device type, browser, host, country searched for, number of pre-experiment searches and the number of search results returned. Using individual impressions, we are also able to calculate the general price level yielded by a given search query by computing the average floor price across all apartments the user has browsed. For the treatment group, we reduce this average by 10% to give the prices users would have seen at baseline. This de-biased price level can also be treated a pre-treatment covariate (in the sense that it is unaffected by treatment status and depends on user-level unobservables such as any specific search terms or filters they may have applied). We then divide this price level into quantiles, which enters subsequent regressions as a set of dummy variables.

Next, we do a balance check on these pre-treatment user-level covariates. In addition to those defined above, we also use the year of the experiment (2019 vs 2020) and the day of week for a user’s first visit for balance checks. We find no significant differences across the treatment and baseline conditions across most of these variables (Figure 3-4). The distributions of four covariates (device type, country searched for, price quantiles and host) appear to be significantly different based on a chi-squared test (Figure 3-5). However, chance imbalances are more likely given the large sample size of our experiment. For these covariates, we calculate the Cramer’s V (Agresti, 1996), which provides a measure of the effect size for a chi-squared test, and find that the size of these differences is negligible - the usual rule of thumb for a small Cramer’s V is 0.1, whereas the values we find are about 100 times (or more) lower than that (Table 3.2). Further, we control for all baseline covariates in our regression estimates to mitigate any remaining concerns of selection on observables.

3.5 Results

3.5.1 ATE of high floor prices

First, we demonstrate the effect of displaying higher “From” prices at the user level. Given random assignment, our goal is to simply estimate the average treatment effect (ATE), i.e, differences in outcomes for the treated group relative to the control group: $E[O_i | \text{Treated}_i = 1] - E[O_i | \text{Treated}_i = 0]$, where O_i is the outcome for user i , and Treated_i indicates their treatment status. Specifically, for each of the 6 outcomes in Table 3.1, we estimate OLS regressions of the following form:

$$\begin{aligned} O_i &= \alpha + \beta \text{Treated}_i + \epsilon_i \\ O_i &= \alpha + \beta \text{Treated}_i + \gamma X_i + \epsilon_i \end{aligned} \tag{3.3}$$

where X_i represents the following covariates: device type, channel type, region

country, number of previous searches, an indicator for the year of the experiment (2019 vs 2020), browser family and host.

The main results are in Table 3.3. We find that treatment group users have fewer outbound clicks, fewer overall searches, fewer dated searches (including a lower likelihood of entering dates), and spend lesser time on the website, indicating a higher bounce rate. The addition of control variables leads to very similar point estimates, reported in Table 3.4. The effect size ranges from 0.5% to 1.3% of the mean outcome values. Although small, these effect sizes are comparable to those obtained in studies of online marketplaces (e.g ad experiments). Furthermore, our intervention is fairly light, since it only changes displayed floor prices by 10% - a stronger (but still realistic) manipulation is likely to yield larger effects.

Consistent with Blake et al. (2018), we also find directional evidence of a negative impact on the likelihood of booking, as well as the total amount spent. As described in the Data section, we do not have enough power to detect statistically significant effects on these sparse outcomes, but the directional results are suggestive of negative impacts across the purchase funnel.⁴

3.5.2 ATE for users exposed to actual prices

We demonstrate that raising floor ARPs has a negative effect on engagement metrics across different stages of the purchase funnel. Next, we want to understand whether the magnitude of discrepancy between floor prices and actual prices might affect outcomes. This can serve as a direct test for the sticker shock hypothesis, which posits that consumers should be averse to a larger gap between true and expected or promised prices. To do this, we narrow down to the set of users ($n=3,041,708$) who conducted at least one search with dates, and thus were exposed to the ‘true’ price

⁴For the same reasons, we are not able to fully explore the “intensive margin” effects shown by Blake et al. (2018).

levels. We can then measure the average difference in dated and undated prices on a per-user basis (henceforth referred to as the wedge, computed at the user level as $\text{AvgPrice}_{dated} - \text{AvgPrice}_{floor}$, where AvgPrice is the average price across all impressions the user has been exposed to).

First, we examine whether the wedge is indeed different across the two groups. Doing so, in Table 3.5, we find that consistent with our treatment assignment, the average difference between dated and undated prices is 10% higher in the control group relative to the treatment group. This demonstrates that displayed prices are on average more accurate (closer to dated prices) in the treatment condition than at baseline. If the sticker shock theory were applicable, we would thus expect a positive treatment effect for this sample.

It is important to note however, that any analysis on this sample is not entirely free from selection bias. This is because, the set of users who choose to enter dates after being faced with a high floor price may differ from those that enter dates in the baseline condition, since the probability of conducting a dated search is lower by about 0.4% in the treatment condition.

While we cannot fully mitigate this bias, a few points are worth noting. Firstly, we do not find evidence of selection on observables by using our pre-treatment covariates and computing propensity scores across treatment and control group users (all values are centred at 0.5 and overlapping). However, selection on unobservables is still possible. The most intuitive source of unobservable selection is that users who enter dates in the treatment condition are less price sensitive. This would mean that users in the treatment group are less likely to be adversely affected by seeing higher prices, and any effects on engagement metrics that we estimate based on this sample would represent a lower bound of the possible negative impact of raising floor prices. Indeed, estimating Equation 3.3 on this sample of users, we continue to find a negative and

significant effect on outbound clicks and dated searches, although the other outcomes are not significant but directionally negative (Table 3.6). Even though the results are more noisy, they do provide evidence against the sticker shock theory since the sign does not flip on any of our coefficients.

Next, we directly look at the wedge as an independent variable, instead of the binary treatment indicator. This will give us a sense of how the magnitude of price difference affects engagement metrics. However, for this analysis, we would ideally wish to vary the wedge exogenously for every user. In other words, we would need to ensure that the average dated prices seen by users in both conditions remain identical. Consistent with the presence of some selection effects, this turns out to be not the case in our sample - we see that dated prices tend to be higher in treatment, and this difference persists even after controlling for other covariates (Table 3.7). We cannot explicitly control for dated prices in Equation 3.3, since it is a post-treatment outcome. Thus, the OLS interpretation of this specification is complicated by the fact that dated prices differ across treatment conditions.

To overcome this concern, we rely on a strategy which generalises the LATE (Local Average Treatment Effect) framework. The idea behind the LATE framework is to characterize compliers among a population of treated individuals and compute the complier average causal effect (CACE) within this population. For instance, if participants are randomized to receive a flyer encouraging them to get vaccinated, the LATE framework allows us to compute the effect of the vaccine on those who actually choose to get vaccinated as a result of the flyer (Angrist and Pischke, 2008).

This example considers that both the endogenous variable (getting vaccinated) and the instrument (receiving a flyer) are binary. The LATE framework can be extended even to the case of continuous endogenous variables (which for us is the wedge). Doing so, we can isolate the part of the variation in the wedge that is due

to treatment, and thus quantify its impact on outcomes. It has been shown that, given mild regularity assumptions, IV independence assumptions identify a weighted average of per-unit causal effects along the length of an appropriately defined causal response function. Conventional instrumental variables and Two-Stage Least Squares procedures can be interpreted as estimating the average causal response to the variable treatment (Angrist and Imbens, 1995).

We implement this strategy by estimating the following regression:

$$O_i = \alpha_i + \beta_1 \cdot \widehat{\text{Wedge}}_i + X_i' \cdot \beta_2 + \tilde{\epsilon}_i, \quad (3.4)$$

$$\text{Wedge}_i = \tilde{\alpha}_i + \gamma_1 \cdot \text{Treated}_i + X_i' \cdot \gamma_2 + \tilde{u}_i. \quad (3.5)$$

The exclusion restriction assumes that the only channel of influence the treatment variable has on outcomes is through its effect on the wedge. Since our treatment is essentially to vary the displayed price, this assumption seems reasonable.

OLS estimates of Equation 3.4 are depicted in Table 3.8. With the OLS estimates, we find a negative and significant relationship between the wedge and outbound clicks, which is suggestive of the sticker shock hypothesis: the larger the wedge, the lesser the number of clickouts. The same pattern holds for time spent on the website. However, for the search and booking value metrics, we find a positive effect, indicating that larger wedges lead to more searches and higher booking value. As mentioned before, these OLS estimates might not lend themselves to a causal interpretation because the wedge is not randomly assigned (average dated prices vary across treatment and control groups). To mitigate these concerns, we move onto the ACR estimation using treatment assignment as an instrument. These estimates are in Table 3.9. As also shown in Table 3.5, we find a strong first stage, indicating that the wedge in the treatment group is on average lower (i.e floor price estimates are more accurate), as

expected. Now, all our other results are consistent with the main specification: all engagement metrics other than dated searches are affected positively, in proportion to the size of the wedge. For example, the coefficient for outbound clicks (0.002) can be interpreted as follows: if the value of the wedge goes up by \$10, the number of outbound clicks increases by 0.02. This provides further support against the sticker shock hypothesis, and in favor of obfuscation - users do not increase their engagement if the wedge is reduced.

3.5.3 Treatment effect heterogeneity: CATE

So far, we have demonstrated that on average, raising floor prices leads to a decline in engagement outcomes across users. This effect persists even within users who are directly exposed to the difference in dated (i.e, ‘true’) and floor prices, further supporting the hypothesis that price salience in the form of more transparent price signals can in fact negatively affect outcomes. In this section, we report some explorations of treatment effect heterogeneity to show which specific groups of users might be affected most strongly. The ATE measures the effect of the intervention over the entire population, but to measure how treatment effects vary across respondent characteristics we estimate a conditional average treatment effect (CATE) (e.g Imai et al. (2013)).

Let $Y_i(1)(Y_i(0))$ denote the potential outcome if individual i is allocated to the treatment (control) group. The causal effect of a treatment on individual i is therefore $Y_i(1) - Y_i(0)$. The average treatment effect (ATE) is then $E[Y_i(1) - Y_i(0)]$. The CATE, measures the average treatment effect for respondents who share a set of characteristics. To formalize this definition, suppose that for each individual we collect J covariates ($j = 1, 2, \dots, J$), $X_i = (X_{i1}, X_{i2}, \dots, X_{iJ})$, with values of the covariates collected in the set χ . We can then define the CATE for covariate profile $x \in \chi$ as $\theta(x)$,

$$\theta(x) = E[Y_i(1) - Y_i(0)|X = x] \quad (3.6)$$

A treatment effect is heterogeneous if the value of Equation 3.6 varies as we consider different strata of participants. As with ATE, random assignment to treatment conditions is sufficient to identify the CATE. To elaborate, the CATE is identified under unconfoundedness, i.e. $Y_i(1), Y_i(0) \perp T_i | X_i$, and overlap, i.e. $0 < \Pr(T_i = 1 | X_i = x) < 1 \forall x$, where T_i denotes the treatment indicator variable.

A CATE estimate can be obtained from a linear model by including interactions between the treatment indicators and the conditioning variable(s) of interest. The inclusion of interaction terms in a linear model is a common technique for exploring the heterogeneity of treatment effects in areas ranging from biomedical science to the social sciences. However, since we have no clear a priori hypotheses about the direction of CATEs for different groups, we explore advances in machine learning to learn these effects instead of reporting results from OLS estimation. We make use of the doubly robust estimator proposed by Foster and Syrgkanis (2019) and others.⁵ This approach flexibly applies machine learning models to create individual level estimates of treatment effects. These estimates can then be projected onto the space of covariates for which we wish to model heterogeneity (while marginalising over all other covariates).

The problem is essentially reduced to the following tasks:

1. predicting the outcome from the treatment and controls
2. predicting the treatment from the controls

⁵Other recent methods to explore heterogeneity such as the causal forest proposed by Wager and Athey (2018) or the BLP method proposed by Chernozhukov et al. (2017) are infeasible in our setting because of our large sample size. These methods are better suited to ‘large p small n’ problems, wherein the number of observations are small relative to the number of covariates and their interactions. In fact, almost all empirical applications of these methods that we have come across deal with sample sizes of less than 1 million.

3. combining these two predictive models in a final stage estimation so as to create a model of the heterogeneous treatment effect.

The approach allows for arbitrary machine learning algorithms to be used for the two predictive tasks, while maintaining many favorable statistical properties related to the final model (e.g. small mean squared error, asymptotic normality, construction of confidence intervals). The latter favorable statistical properties hold if either the first or the second of the two predictive tasks achieves small mean squared error (hence the name doubly robust).

The final stage regression estimated through this method is meaningful even if the space of functions over which we minimize the final regression loss does not contain the true CATE function. In that case, the method will estimate the projection of the CATE function onto the space of models over which we optimize in the final regression. For instance, this allows one to perform inference on the best linear projection of the CATE function or to perform inference on the best CATE function on a subset of features that could potentially be creating heterogeneity, without making any further assumptions on how that heterogeneity looks like (Foster and Syrgkanis, 2019). For implementational details, please refer to Section B.1.

In summary, this procedure yields individual level treatment effect estimates which can be projected onto different sets of covariates to provide estimates of CATEs (while controlling for baseline differences across all other covariates). First, we estimate a projection of the CATEs onto a constant. This is akin to estimating the ATE from the individual treatment effect estimates. Comparing these estimates with the OLS estimates in Table 3.4, we find very consistent results, thus lending some credibility to our estimate of individual level effects (Figure 3.7).

Next, we project these estimates, in turn, onto our pre-treatment covariates. We don't find evidence of very large heterogeneous effects - across various covariates,

effects continue to be negative and not very different from each other. We feel that this ‘null result’ is also relevant from a policy perspective, since it indicates that based on the given covariates, we don’t observe consistently positive and significant effects for any user group which would justify rolling out the high floor price treatment to that group. Nonetheless, we report results from the two sets of covariates (acquisition channel and price level) that yielded the most interesting patterns (other results are available in Section B.2):

1. Users exposed to more expensive listings tend to click out less in treatment (-0.04) relative to those exposed to less expensive listings (-0.015). This is interesting because we would expect users searching for more expensive listings to be less price sensitive and thus less affected by higher displayed prices (Figure 3-8).
2. Users searching for less expensive listings (lower price levels) have a significantly lower propensity to book, as well as reduced booking value in treatment. On the other hand, users searching for more expensive listings reduce their engagement (time on site, number of searches) more in treatment relative to the former group. This indicates that higher floor prices might have differential negative effects across consumers of different price sensitivities, but does not have a positive and significant impact on any group (Figure 3-8).
3. Users acquired via organic Google search and Facebook tend to exhibit some positive effects - in particular, these users spend significantly more time on the website. Facebook users also tend to have a higher propensity to book and higher spending. Although the 95% confidence interval overlaps 0, these estimates are more precisely estimated than those of the other categories (Figure 3-9). Unfortunately we do not have enough data to investigate this further,

but it does provide preliminary evidence that acquisition channel might have an effect on user perceptions of floor prices. Future work can aim to investigate this effect in greater detail.

3.6 Conclusion

Advertised reference prices (ARPs) are an important tool with which firms can calibrate the price expectations of consumers. ARPs have been shown to have an impact on consumer evaluations and purchase behavior in a variety of domains. However, most previous applications have looked at ARPs that are framed as discounts and serve to emphasize transaction utility. In this paper, we experimentally explore a relatively understudied variant of ARPs, namely the “Starting from” (floor) price commonly displayed by vacation rental platforms. Using results from two large-scale field experiments conducted on Holidu.com, we find that higher floor prices tend to decrease engagement metrics at all levels of the purchase funnel: namely, they lead to fewer outbound clicks, fewer searches as well as lesser time spent on the website. We also find noisy but suggestive evidence of a negative effect on booking level outcomes such as the propensity to book as well as the total booking amount.

Our findings are supportive of how firms can benefit (at least in the short term) by engaging in price obfuscation and reducing salience. This view is consistent with several recent papers examining price salience in digital markets (most notably Blake et al. (2018)). However, they seemingly contradict the sticker shock theory, which posits that consumers react adversely if there is a large discrepancy between expected and true prices. Although raising floor prices amounts to reducing this discrepancy, consumers react adversely even when they are exposed to true (i.e. dated) prices on the platform. Finally, we check for treatment effect heterogeneity and do not find large heterogeneity in these effects. Nevertheless, we find some evidence that the channel

of acquisition and the general price levels of the listings searched for can have some differential effects on outcomes. In particular, we find that users searching for more expensive listings reduce their clicks the most when faced with higher floor prices, whereas users searching for cheaper listings tend to exhibit a greater negative effect on booking level outcomes. In addition, users acquired through Facebook tend to spend more time on the website, and also exhibit greater booking rates and spending when exposed to higher floor prices. We do not have enough data to clearly delineate the underlying mechanism for these observations, but they open up interesting avenues for future research.

Our results suggest that, in a bid to provide more accurate and transparent price estimates, firms may implicitly end up adversely affecting engagement outcomes. It is thus worth carefully considering how these transparency signals may be conveyed in an online environment where users have limited attention. Our data from Holidu.com encompasses users searching for accommodation from across the globe, and can thus generalize to other global vacation rental or aggregation platforms. That being said, the extent of the effect as well as other moderating factors are all empirical questions worth examining on different platforms and domains. In particular, online marketplaces exhibit many other forms of less conventional ARPs. We have focused our attention to floor prices, but average prices and price ranges are also commonly seen. It is worth examining in the future how these signals might affect consumer perceptions.

Another limitation of our study is that we have manipulated prices only by 10%. It is thus worth understanding what the limits are to the effects we observe. Perhaps one moderating factor could be whether the prices are believable (as shown by Ngwe (2018)), and low floor prices continue to attract consumers as long as they are not unrealistically low ("Starting from \$1"). The optimal 'threshold' is an empirical

question to investigate in future work.

Finally, we do not study long term effects. It would be worth tracking the same users over time to see if repeated exposure to low (vs high) floor prices has an effect. The ability to track users long term would also enable the collection of more covariates that can enrich the heterogeneity analysis described above. Unfortunately this is infeasible in our current setting since the majority of users are one-off and do not have an account with the website.

On the surface, our findings indicate that consumers do not value accurate or transparent prices, which is somewhat surprising. However, part of the effect may arise from how floor price information is presented to consumers. One possible mechanism to ‘debias’ consumer perceptions could perhaps be displaying all prices on a calendar, rather than providing dated prices only on request. This takes into account consumer inattention by making the difference between floor prices and dated prices more direct, while also promoting greater price transparency. As an initial test of this hypothesis, we examined the effect of floor price changes on Airbnb.com using data collected in 2014-2015 (described in Section B.3). At the time, users could see prices for every calendar date (like Google Flights) along with a ‘From’ price displayed at the top, hence arguably, comparisons between actual and floor prices could easily be made. Interestingly, this data indicates (1) a positive correlation between outcomes and floor price levels, and (2) a negative correlation between outcomes and the wedge. Hence, these results lean towards supporting the sticker shock hypothesis. However, given the observational nature of this data and the lack of a credible exogenous shock in floor prices, this is preliminary evidence that should be experimentally investigated.

Our study is among the first to study ARPs using field experiments that explicitly manipulate displayed prices. Doing so, we capture how floor prices affect user engagement on a large vacation rental platform and provide policy-relevant findings

that can benefit platforms and consumers.

3.7 Tables

Table 3.1: Summary statistics of our main outcomes of interest.

	Num. users	Mean	Std	Max	Min
Number of searches	6,385,717	3.340	6.209	1,031	0
Outbounds	6,385,717	2.028	4.809	537	0
Dated searches	6,385,717	2.330	4.923	857	0
Booked	6,385,717	0.005	0.072	1	0
Booking value	6,385,717	4.869	110.091	45,855.750	0
Time on Site (hours)	6,385,717	30.075	113.053	1,157.147	0

Table 3.2: Cramer's V computed for covariates that had a statistically significant different in balance checks.

	Cramer's_V	Covariate
1	0.0007	Device Type
2	0.0035	Price
3	0.0006	Country
4	0.0005	Host

Table 3.3: The effect of raising floor prices on user-levels outcomes: not including pre-treatment covariates.

	<i>Dependent variable:</i>					
	Outbound	Dated Searches	All searches	Booked	BookingValue	Time(in hrs)
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	-0.027*** (0.004)	-0.012*** (0.004)	-0.012** (0.005)	-0.0001 (0.0001)	-0.052 (0.087)	-0.179** (0.089)
Constant	2.041*** (0.003)	2.336*** (0.003)	3.345*** (0.003)	0.005*** (0.00004)	4.895*** (0.062)	30.165*** (0.063)
Pre-treatment covariates	No	No	No	No	No	No
Observations	6,385,717	6,385,717	6,385,717	6,385,717	6,385,717	6,385,717
Adjusted R ²	0.00001	0.00000	0.00000	0.00000	-0.00000	0.00000

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3.4: The effect of raising floor prices on user-levels outcomes: including pre-treatment covariates.

	<i>Dependent variable:</i>					
	Outbound	Dated Searches	All searches	Booked	Booking Value	Time (in hrs)
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	-0.028*** (0.004)	-0.012*** (0.004)	-0.012** (0.005)	-0.0001 (0.0001)	-0.057 (0.087)	-0.169* (0.088)
Constant	0.699*** (0.012)	1.772*** (0.012)	2.241*** (0.015)	0.003*** (0.0002)	2.958*** (0.268)	-4.937*** (0.270)
Pre-treatment covariates	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,385,717	6,385,717	6,385,717	6,385,717	6,385,717	6,385,717
Adjusted R ²	0.024	0.033	0.037	0.003	0.001	0.036

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3.5: The average user-level difference in dated and floor prices (Wedge) as a function of treatment status.

	<i>Dependent variable:</i>
	Wedge
Treated	-7.073*** (0.279)
Constant	167.300*** (0.203)
Observations	3,019,877
R ²	0.0002
Adjusted R ²	0.0002
Residual Std. Error	242.218 (df = 3019875)
F Statistic	641.597*** (df = 1; 3019875)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Table 3.6: The impact of raising floor prices on users who have been exposed to dated prices.

	<i>Dependent variable:</i>					
	Outbound (1)	Dated Searches (2)	All searches (3)	Booked (4)	Booking Value (5)	Time (in hrs) (6)
Treated	-0.017** (0.007)	-0.013* (0.007)	-0.010 (0.009)	-0.0001 (0.0001)	-0.005 (0.170)	-0.059 (0.153)
Constant	1.443*** (0.094)	4.960*** (0.094)	5.594*** (0.118)	0.012*** (0.002)	6.076*** (2.292)	-7.954*** (2.069)
Pre-treatment covariates	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,041,708	3,041,708	3,041,708	3,041,708	3,041,708	3,041,708
Adjusted R ²	0.031	0.034	0.038	0.004	0.002	0.048

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3.7: The prices seen by users when they enter dates as a function of treatment status.

	<i>Dependent variable:</i>	
	Average Dated Price	
	(1)	(2)
Treated	33.788** (15.177)	33.863** (15.176)
Constant	267.508*** (10.724)	339.903* (203.751)
Pre-treatment covariates	No	Yes
Observations	3,041,708	3,041,708
Adjusted R ²	0.00000	0.0001

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

Table 3.8: The impact of the wedge on outcomes on interest.

	<i>Dependent variable:</i>					
	Outbound (1)	Dated Searches (2)	All searches (3)	Booked (4)	Booking Value (5)	Time (in hrs) (6)
Wedge	-0.001*** (0.00002)	0.001*** (0.00002)	0.001*** (0.00002)	-0.00000*** (0.00000)	0.009*** (0.0004)	-0.010*** (0.0004)
Constant	1.497*** (0.094)	4.856*** (0.094)	5.497*** (0.118)	0.012*** (0.002)	5.283** (2.290)	-7.080*** (2.068)
Pre-treatment covariates	Yes	Yes	Yes	Yes	Yes	
Observations	3,041,708	3,041,708	3,041,708	3,041,708	3,041,708	3,041,708
Adjusted R ²	0.031	0.035	0.039	0.004	0.002	0.048

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3.9: The impact of the wedge on outcomes of interest, instrumented by treatment status.

<i>Dependent variable:</i>							
	Outbound	Dated Searches	All searches	Bookings	Booking Value	Time (in hrs)	Stage1
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Wedge	0.002** (0.001)	0.00001 (0.001)	0.003*** (0.001)	0.00000 (0.00001)	0.019 (0.021)	0.042** (0.019)	
Treated							-8.162*** (0.244)
Constant	0.909*** (0.110)	4.709*** (0.110)	4.596*** (0.137)	0.011*** (0.002)	3.367 (2.662)	-17.007*** (2.412)	68.454*** (3.284)
Pre-treatment covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,041,708	3,041,708	3,041,708	3,041,708	3,041,708	3,041,708	3,041,708
Adjusted R ²	0.022	0.033	0.036	0.004	0.002	0.042	0.040

Note:

*p<0.1; **p<0.05; ***p<0.01

3.8 Figures

www.tripadvisor.com > ... > Boston ⋮
THE 10 BEST Hotels in Boston, MA for 2021 (from \$90 ...
The #1 Best Value of 244 places to stay in **Boston**. Restaurant. Room service. Special offer.
Hotel website. **Boston** Park Plaza. Show Prices. #2 Best Value of 244 ...

www.booking.com > ... > Hotels in Massachusetts ⋮
The 10 Best Boston Hotels (From \$67) - Booking.com
Hotels located in the center of **Boston**. Four Seasons **Hotel** One Dalton Street, **Boston**. **Hotel**
in Back Bay, **Boston**. Hyatt Centric Faneuil Hall **Boston**. **Hotel** in Financial District, **Boston**.
Courtyard **Boston** Downtown/North Station. **Hotel** in Downtown **Boston**, **Boston**. Loews
Boston Hotel. AC **Hotel** by Marriott **Boston** Cleveland ...

www.expedia.com > ... > Massachusetts ⋮
Top Hotels in Boston, MA from \$86 (FREE cancellation on ...
Check **Boston hotel** prices · The Revolution **Hotel** · Hyatt Regency **Boston** Harbor · Embassy
Suites **Boston** Logan Airport · You could be seeing lower prices · Hilton ...

Figure 3.1: Example of advertised floor prices appearing in a Google search for ‘Hotels in Boston.’

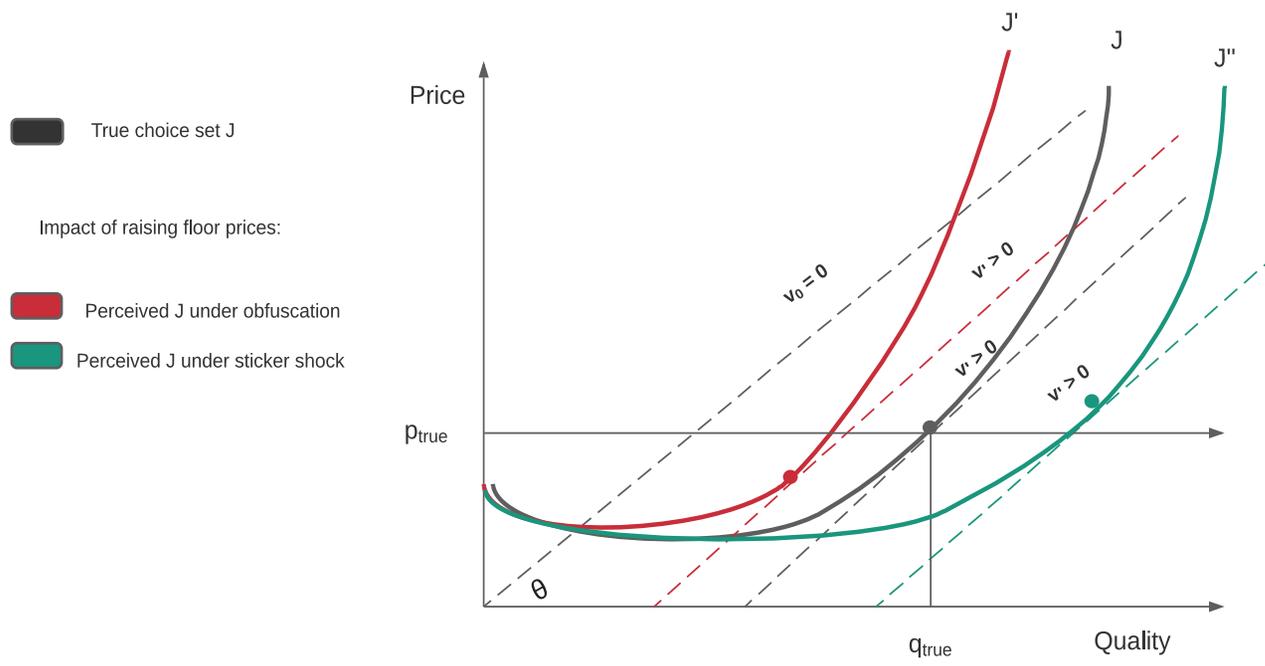


Figure 3·2: A stylised model to depict the possible impacts of raising floor prices.

Spain > Balearic Islands > Majorca

Majorca: 12,663 accommodations found Sort by

Enter your travel dates to see available accommodations!

Good deal

Holiday Apartment La Cabanya ...
 Capdepera
 2 pers. 1 bedroom 753ft²
 250yd to the beach
 4.5 ★★★★★ (21 reviews)

from **\$42** per night

More details > **VIEW OFFER**

Good deal

SON SABATER (ES MOLINO)
 Sa Pobla
 4 pers. 4 bedrooms 1,292ft²
 4.4mi to the beach
 4.6 ★★★★★ (14 reviews)

from **\$92** per night

More details > **VIEW OFFER**

Good deal

CAN SETI
 Alaró
 3 pers. 1 bedroom 861ft²
 4.9 ★★★★★ (7 reviews)

from **\$90** per night

More details > **VIEW OFFER**

Good deal

Apartment Close to the Beach w...
 Port de Pollença
 4 pers. 2 bedrooms 753ft²
 500yd to the beach
 4.6 ★★★★★ (1 review)

from **\$64** per night

More details > **VIEW OFFER**

[View map](#)

Reset all filters

Property type

- Vacation house
- Villa
- Vacation apartment
- Condo
- Cabins
- Beach home
- Cottage
- Ski Lodge
- Hotel

Price

Drag the handles to select your price range

\$0 - \$500+ per night

Book on **holidu**

- Fast and secure booking directly on Holidu

Amenities

- Pool

Spain > Balearic Islands > Majorca

Majorca: 12,663 accommodations found Sort by

Good deal

Holiday Apartment La Cabanya ...
 Capdepera
 2 pers. 1 bedroom 753ft²
 250yd to the beach
 4.5 ★★★★★ (21 reviews)

from **\$46** per night

More details > **VIEW OFFER**

Good deal

SON SABATER (ES MOLINO)
 Sa Pobla
 4 pers. 4 bedrooms 1,292ft²
 4.4mi to the beach
 4.6 ★★★★★ (14 reviews)

from **\$101** per night

More details > **VIEW OFFER**

Good deal

CAN SETI
 Alaró
 3 pers. 1 bedroom 861ft²
 4.9 ★★★★★ (7 reviews)

from **\$99** per night

More details > **VIEW OFFER**

Good deal

Apartment Close to the Beach w...
 Port de Pollença
 4 pers. 2 bedrooms 753ft²
 500yd to the beach
 4.6 ★★★★★ (1 review)

from **\$70** per night

More details > **VIEW OFFER**

Figure 3.3: Example of experimental manipulation on Holidu.com.

Characteristic	N	0, N = 3,192,202 ¹	1, N = 3,193,515 ¹	p-value ²
channel	6,385,717			>0.9
BING_PAID_SEARCH		162,292.000 (5.084%)	162,968.000 (5.103%)	
DIRECT		121,754.000 (3.814%)	121,884.000 (3.817%)	
FACEBOOK_PAID_SEARCH		70,088.000 (2.196%)	69,959.000 (2.191%)	
GOOGLE_DISPLAY		54,128.000 (1.696%)	54,042.000 (1.692%)	
GOOGLE_ORGANIC		47,331.000 (1.483%)	47,339.000 (1.482%)	
GOOGLE_PAID_SEARCH		2,596,574.000 (81.341%)	2,597,272.000 (81.330%)	
Misc		74,399.000 (2.331%)	74,200.000 (2.323%)	
REFERRER		65,636.000 (2.056%)	65,851.000 (2.062%)	
browserfamily	6,385,717			0.12
Chrome		1,925,608.000 (60.322%)	1,925,196.000 (60.285%)	
Firefox		248,567.000 (7.787%)	249,344.000 (7.808%)	
Internet Explorer		111,595.000 (3.496%)	112,280.000 (3.516%)	
Microsoft Edge		155,675.000 (4.877%)	156,728.000 (4.908%)	
Misc		39,750.000 (1.245%)	39,353.000 (1.232%)	
Safari		711,007.000 (22.273%)	710,614.000 (22.252%)	
numberofsearchinjourney	6,385,717	2.590 (9.937)	2.592 (9.784)	0.8
countofresults	6,385,717	16,863.960 (46,835.771)	16,963.393 (47,064.518)	0.15
experiment_type	6,385,717			0.7
2019		593,414.000 (18.589%)	593,269.000 (18.577%)	
2020		2,598,788.000 (81.411%)	2,600,246.000 (81.423%)	
day	6,385,717			0.3
1		602,639.000 (18.878%)	603,377.000 (18.894%)	
2		484,156.000 (15.167%)	484,959.000 (15.186%)	
3		436,535.000 (13.675%)	435,099.000 (13.624%)	
4		449,845.000 (14.092%)	449,728.000 (14.083%)	
5		445,656.000 (13.961%)	446,964.000 (13.996%)	
6		374,815.000 (11.742%)	373,941.000 (11.709%)	
7		398,556.000 (12.485%)	399,447.000 (12.508%)	

¹n (%); Mean (SD)

²Pearson's Chi-squared test; Wilcoxon rank sum test

Figure 3-4: Distribution of pre-treatment covariates across treatment conditions: no significant differences.

Characteristic	N	0, N = 3,192,202 ¹	1, N = 3,193,515 ¹	p-value ²
cut_price	6,107,993			<0.001
(24,51]		749,520.000 (24.649%)	758,468.000 (24.728%)	
(51,68]		774,952.000 (25.486%)	773,828.000 (25.229%)	
(68,96]		765,085.000 (25.161%)	762,671.000 (24.865%)	
(96,392]		751,174.000 (24.704%)	772,295.000 (25.179%)	
Unknown		151,471	126,253	
regioncountry	6,385,717			0.011
Australia		34,698.000 (1.087%)	34,311.000 (1.074%)	
Austria		82,903.000 (2.597%)	82,752.000 (2.591%)	
Belgium		34,624.000 (1.085%)	34,628.000 (1.084%)	
Brazil		69,742.000 (2.185%)	70,859.000 (2.219%)	
Croatia		56,406.000 (1.767%)	57,145.000 (1.789%)	
France		398,069.000 (12.470%)	397,810.000 (12.457%)	
Germany		471,044.000 (14.756%)	472,510.000 (14.796%)	
Italy		276,483.000 (8.661%)	276,652.000 (8.663%)	
Misc		232,470.000 (7.282%)	231,871.000 (7.261%)	
None		231,100.000 (7.240%)	232,334.000 (7.275%)	
Poland		48,191.000 (1.510%)	48,362.000 (1.514%)	
Portugal		155,129.000 (4.860%)	155,750.000 (4.877%)	
Spain		677,567.000 (21.226%)	676,235.000 (21.175%)	
The Netherlands		143,500.000 (4.495%)	142,526.000 (4.463%)	
United Kingdom		181,416.000 (5.683%)	181,527.000 (5.684%)	
USA		98,860.000 (3.097%)	98,243.000 (3.076%)	

¹n (%)

²Pearson's Chi-squared test

Figure 3-5: Distribution of pre-treatment covariates across treatment conditions: significant differences according to a Chi-Sq. test.

Characteristic	N	0, N = 3,192,202 ¹	1, N = 3,193,515 ¹	p-value ²
devicetype	6,385,717			0.034
Computer		1,170,597.000 (36.671%)	1,174,010.000 (36.762%)	
Misc		658.000 (0.021%)	720.000 (0.023%)	
Mobile		1,857,521.000 (58.189%)	1,855,548.000 (58.104%)	
Tablet		163,426.000 (5.120%)	163,237.000 (5.112%)	
host	6,385,717			0.049
Misc		242,105.000 (7.584%)	242,362.000 (7.589%)	
www.holidu.be		50,178.000 (1.572%)	50,461.000 (1.580%)	
www.holidu.ch		35,761.000 (1.120%)	36,038.000 (1.128%)	
www.holidu.co.uk		145,185.000 (4.548%)	145,380.000 (4.552%)	
www.holidu.com		75,898.000 (2.378%)	75,682.000 (2.370%)	
www.holidu.com.au		37,531.000 (1.176%)	37,019.000 (1.159%)	
www.holidu.com.br		70,259.000 (2.201%)	71,393.000 (2.236%)	
www.holidu.de		866,243.000 (27.136%)	866,910.000 (27.146%)	
www.holidu.dk		37,568.000 (1.177%)	37,829.000 (1.185%)	
www.holidu.es		418,955.000 (13.124%)	417,874.000 (13.085%)	
www.holidu.fr		424,040.000 (13.284%)	424,387.000 (13.289%)	
www.holidu.it		169,391.000 (5.306%)	169,471.000 (5.307%)	
www.holidu.nl		189,723.000 (5.943%)	188,407.000 (5.900%)	
www.holidu.pt		114,323.000 (3.581%)	114,629.000 (3.589%)	
www.hundredrooms.co.uk		107,670.000 (3.373%)	107,579.000 (3.369%)	
www.hundredrooms.com		84,306.000 (2.641%)	84,005.000 (2.630%)	
www.hundredrooms.de		75,006.000 (2.350%)	75,776.000 (2.373%)	
www.vakantiehuisdirect.nl		48,060.000 (1.506%)	48,313.000 (1.513%)	

¹n (%)

²Pearson's Chi-squared test

Figure 3-6: Distribution of pre-treatment covariates across treatment conditions: significant differences according to a Chi-Sq. test.

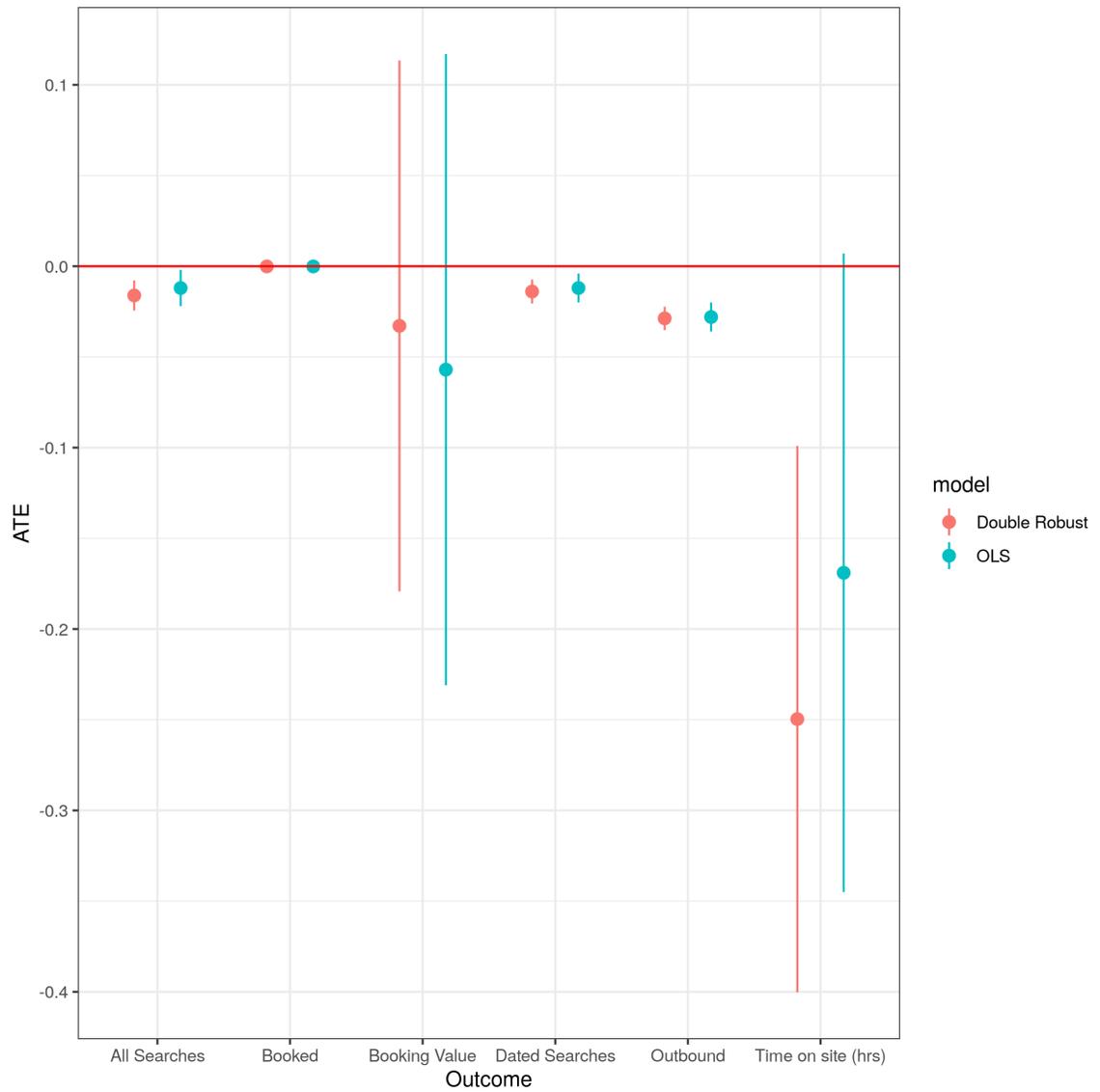


Figure 3-7: Comparison of point estimates and standard errors across OLS and doubly robust estimates. We find very consistent results.

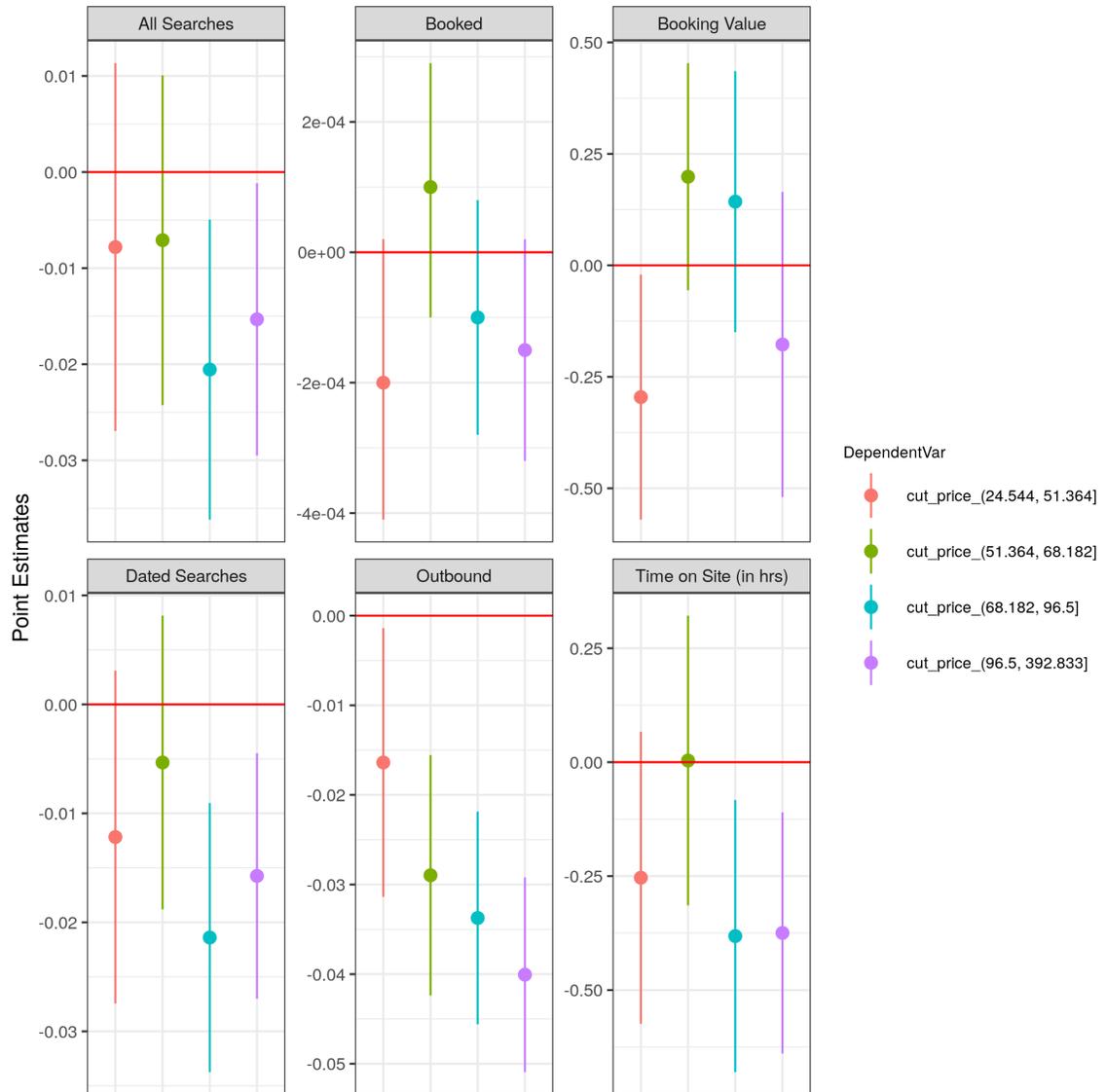


Figure 3-8: Treatment effect heterogeneity across price levels.

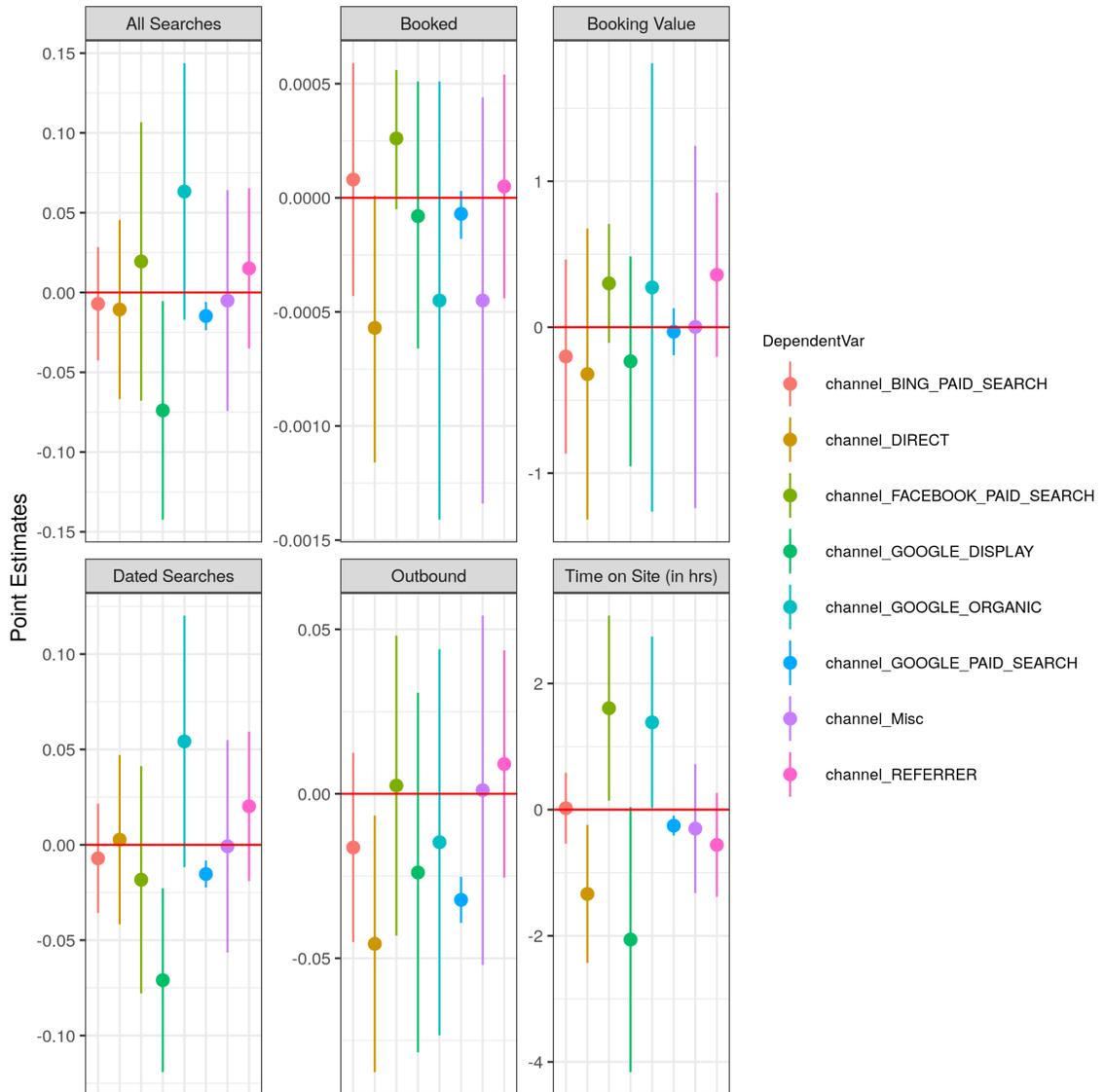


Figure 3-9: Treatment effect heterogeneity across acquisition channel.

Chapter 4

Conclusion

In my PhD research, I try to understand how consumers respond to different product page “cues” that they are exposed to in an online setting, and how these cues might eventually affect the profitability of the platform. To causally identify the effects I estimate, I use an experimental/quasi-experimental framework along with tools from empirical econometrics as well as machine learning.

This research has direct implications for the design choices made by online platforms. In particular, I show that (1) the design choice of including consumer-contributed Q&As can reduce frictions in the matching process of consumers with products and thus lead to higher satisfaction, and (2) not anticipating consumer responses to salient prices can have a detrimental effect on consumer engagement metrics. Building on these ideas, future research can further examine the optimal use and limits of such design choices. My particular setting explores a subset of empirical applications but it will be worthwhile to further investigate the overall welfare implications of these choices.

My research agenda going forward is to continue exploring (1) the appropriate tools that firms can leverage to help consumers make better purchase decisions online; (2) the design of reputation systems, their inherent biases, as well as how they differ based on the type of platform and product; (3) advances in causal machine learning and large scale text analysis/natural language processing that are applicable to the problems described above.

To this end, in ongoing projects that emerge from my dissertation research, I am exploring other related aspects of online decision making. In one project, I am examining specific assumptions made by consumers when they submit a product rating online, and how a simulated generative model can be used to incorporate those assumptions and back out specific “costs” of leaving a review. In particular, we propose a likelihood free inference engine using neural networks (adapted from Tejero-Cantero et al. (2020)) that can infer these costs in a reliable and scalable way. Our model takes only the histogram of observed reviews as input, and therefore can be used to model correlations of inferred cost parameters with various product features. As a preliminary proof of concept, we apply the model to a dataset of 450,000 product reviews of 939 products submitted on Amazon.com. We find that the cost to leaving a negative review is much greater than a positive review (Figure 4.1), and this cost is further correlated with price, brand, and the number of reviews a product has. Gaining a better understanding of the dynamics of reputation systems, namely, the conditions under which ratings are submitted, is crucial for marketers, brand managers, and designers of digital platforms, who can leverage this information to stimulate further reviews and better manage user generated content.

In a second project, I am focusing my attention on Goodreads.com, a major website for book reviews. Goodreads also allows consumer Q&As, and I am currently trying to understand how the role played by Q&As in this setting might differ from that proposed in Chapter 2. Initial results from roughly 63,000 books indicate that Q&As contain a mix of subjective (age appropriateness of a certain book) and objective (where to find a copy) information (Figure 4.2). Further, there is interesting heterogeneity in the kinds of genres that attract a question - for example, young adult, audiobooks and thrillers have a larger proportion of books with at least one question than non-fiction or food and drink (Figure 4.3). We did not specifically explore why

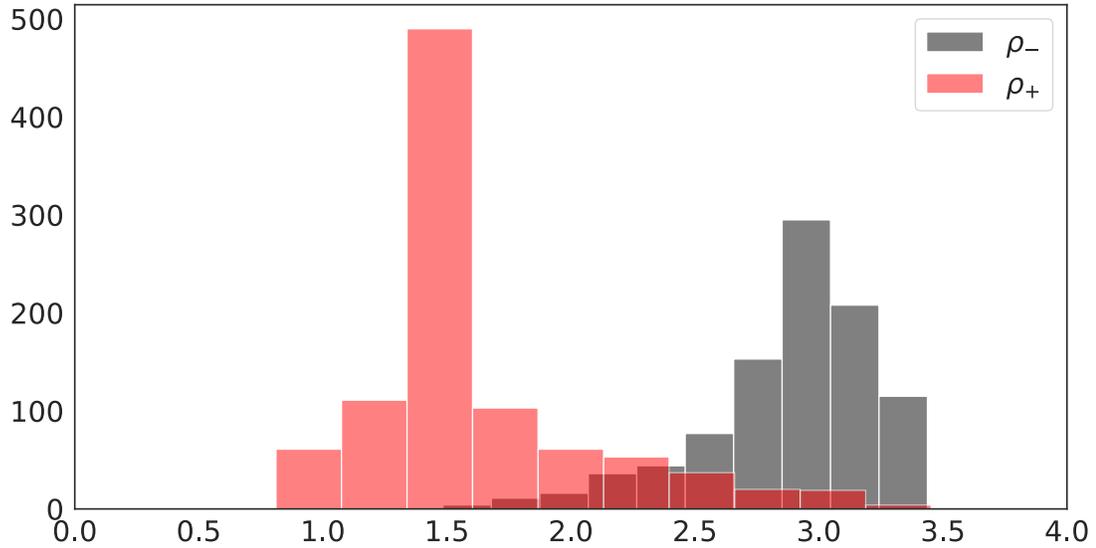


Figure 4.1: The ‘cost’ of leaving a review formulated as a threshold ρ . Our generative model simulates several histograms with different values of ρ_+ and ρ_- - the values here indicate the mean posterior ρ values computed across 939 products on Amazon.com.

certain kinds of products might get questions in Chapter 2, so this research can serve as an interesting complement to our existing results.

In the future, I also want to find more applications of machine learning to enrich causal inference. I started using machine learning to create interesting covariates (like vector representations of text, or new categorical variables) that can then enter regression models and lend more richness and interpretability to estimation. In addition to such applications, I have recently begun to explore several “causal machine learning” tools, which are particularly useful for high dimensional causal analysis. I use one such example (doubly robust inference for heterogeneous treatment effects) in Chapter 3. Causal ML is a burgeoning area that will likely form the foundations for much of causal inference in the years to come, and I hope to find more applications for these tools in my current and future projects. As digitization continues to exert an all-encompassing influence on buyers, sellers and societies alike, it is important

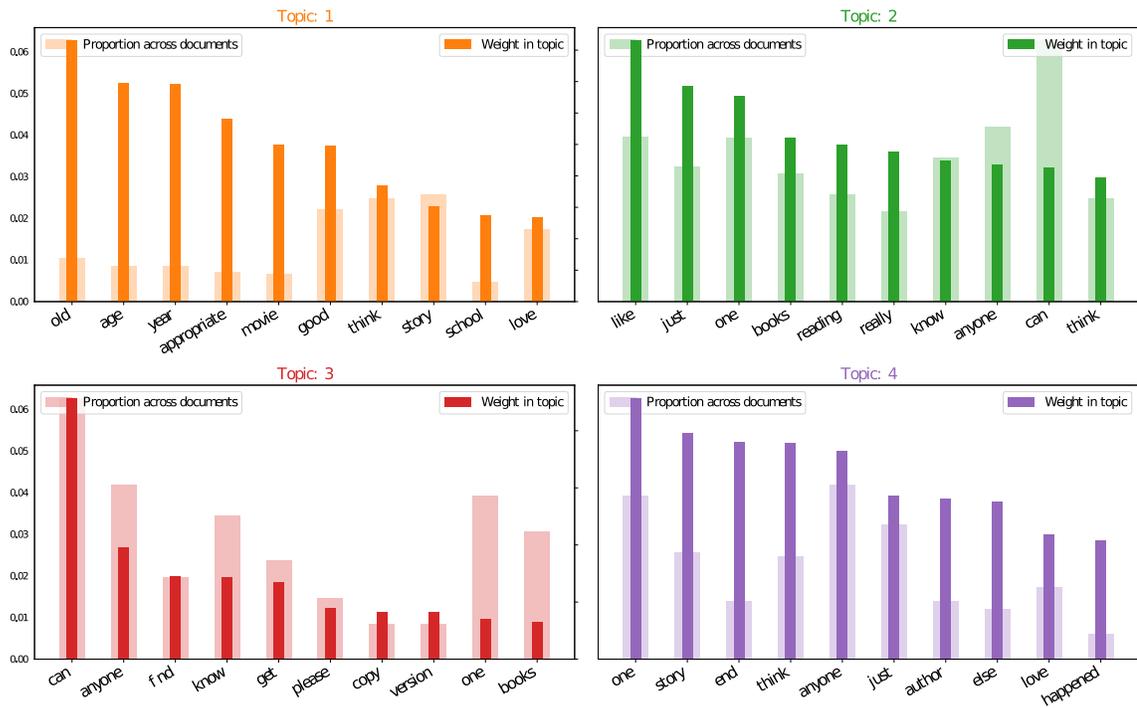


Figure 4.2: Proportion across documents and topic-specific importance of keywords for the top 4 topics obtained by training an LDA topic model on question text across 63,000 books on Goodreads.com.

to continue examining various aspects of the digital revolution and how it impacts stakeholders at large.

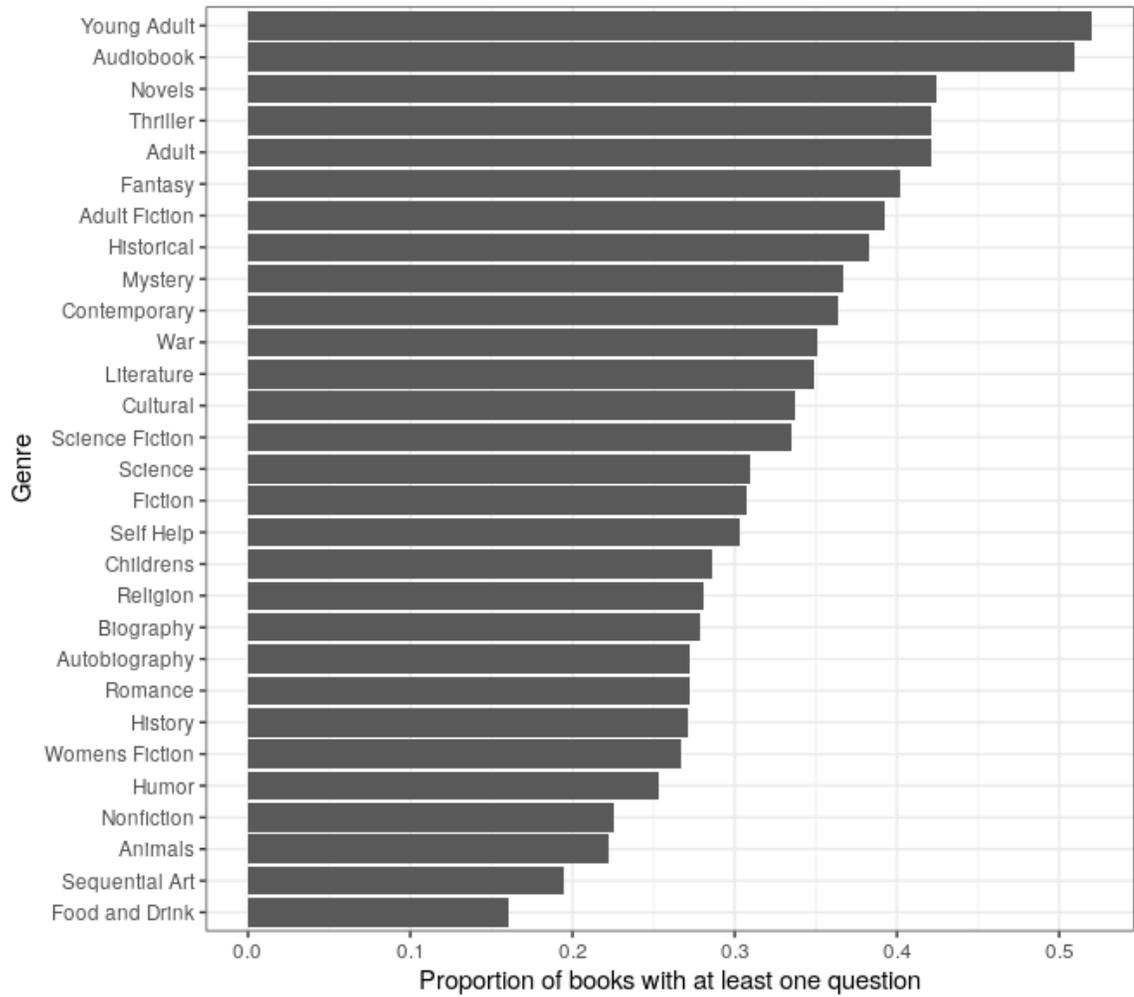


Figure 4.3: The proportion of books within the top 30 genres that receive at least one question. Each genre accounts for at least 1% of books in the sample.

Appendix A

Interacting User Generated Content Technologies: How Questions and Answers Affect Consumer Reviews

A.1 Tables and Figures

Table A.1: LDA topics extracted from the Q&A corpus.

Topic	Highest Probability Words
Dimensions	height width depth dimensions length size
Guarantee/warranty	buy bought guarantee product warranty year
Compatibility (computers)	ipod compatible laptop windows work download
Dimensions (furniture) 1	fit sofa flat size item dimensions
Attributes (kitchen appliance)	oven grill light switch time microwave
Dimensions (furniture) 2	door side drawers left doors open
Attributes (furniture)	weight table chair chairs back seat
Installation queries (kitchen)	cooker gas included installation electric include
Compatibility (phone)	phone card sim work memory phones
Compatibility (home appliance)	box work connect record freeview internet
Attributes (lamps)	glass light pole lid lamp plastic
Attributes (power socket)	cable usb plug battery socket power
Attributes (entertainment system)	work player dvd play remote samsung
Dimensions (bed)	bed mattress size fit base double
Queries (home appliance)	fridge freezer free dryer long wash
Dimension (furniture) 3	wall unit shelves shelf top fit
Instructions (camera/printer)	camera printer print clock ink set
Description discrepancies	product description confirm question correct states
Color/finish (furniture)	colour made white black wood match
Attributes (home appliance)	machine water washing make hot filter

Table A.2: LDA topics extracted from the review corpus.

Topic	Highest Probability Words
Quality (vacuum)	easy great good cleaner clean product
Quality (electronics)	sound good great clock quality set
Quality (phone/camera)	phone easy set good camera features
Quality (home appliance)	kettle water good iron machine toaster
Quality (bedclothes)	bed duvet comfortable warm mattress pillows
Value for money 1	good job easy product money price
Product instructions	put instructions wall screws holes fit
Product returns	store item product service delivery back
Replacement	bought years printer replace buy good
Gifts	bought great room son loves year
Quality (clothes line)	put clothes easy good cover sturdy
Dimensions (storage unit)	storage easy small space put unit
Value for money 2	product money great good recommend excellent
Value for money (negative)	it's bit reviews price cheap buy
Quality (garden tools)	good job cut light easy small mower
Discounts	price good quality great bargain sale
Quality (lighting)	light colour nice lovely lamp room
Assembly instructions	easy good put table assemble money
Quality (kitchen appliance)	easy clean great cooker microwave food
Quality (poor)	quality bin back poor plastic cheap

Table A.3: Review volume following the first answer.

	OLS (Rating proxy)	OLS (Var. proxy)	OLS (Fit proxy)
POST \times Low Rating	-0.037 (0.033)		
POST \times High Variance		0.020 (0.037)	
POST \times Fit			-0.157 (0.105)
POST	0.089* (0.050)	0.074 (0.054)	0.085* (0.047)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes
Observations	37,863	37,863	37,863

Note:

*p<0.1; **p<0.05; ***p<0.01

Standard errors clustered at the product level.

Table A.4: Pageview volume following the first answer.

	OLS (Rating proxy)	OLS (Var. proxy)	OLS (Fit proxy)
POST \times Low Rating	0.318 (0.468)		
POST \times High Variance		0.607 (0.479)	
POST \times Fit			0.710 (0.930)
POST	-0.424 (0.410)	-0.506 (0.408)	-0.349 (0.349)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes
Observations	20,209	20,209	20,209

Note:

*p<0.1; **p<0.05; ***p<0.01

Standard errors clustered at the product level.

(a) Yelp.

(b) Google reviews.

(c) TripAdvisor.

(d) Amazon.

Figure A.1: Q&A technology on different platforms.

Categorization Instructions (PLEASE READ!) (Click to collapse)		
<p>You will read a negative review posted on an online shopping website. You need to indicate why you think the customer was unhappy with the purchase. Based on the following, please indicate what you feel is the main/strongest reason behind the negative review.</p>		
Category	Description	Examples
Poor fit	The product was not what the customer expected (e.g., unclear or wrong product description, does not perform the expected function, not of the expected dimensions/colour)	"What a waste of money. I bought it specifically because I had a friend coming and it is much too small for an adult to sleep on and barely big enough for a child. Absolutely useless."
Poor quality	The product quality was bad (e.g., missing parts, flimsy, damaged after a couple of uses)	"This unit sounds tinney there is no bass the TV sounds better , if this was not a christmas present from the wife it would end up in the loft with the rest of the rubbish"
Other	Store related issues (e.g., delivery delay, return hassles) and miscellaneous	"Reserved two at the Bristol main outlet. Took 4 days to arrive! Could have ordered two from RS at 6pm and collected next day at 10am!"

Figure A.2: Screenshot of survey shown to workers on MTurk.

Table A.5: Examples of reviews indicating fit and quality issues.

Fit issues	Quality issues
<p>1. What a waste of money. I bought it specifically because I had a friend coming and it is much too small for an adult to sleep on and barely big enough for a child. Absolutely useless. (3 stars)</p> <p>2. Bought for an occasional put-u-up for the grandchild on sleep over - suitable for small child, folding away to make a convenient bed chair. NOT SUITABLE FOR 10+. As it's close to the ground and contains no metal or sharp pieces, the child cannot hurt itself if rolls out of bed!! For what it is could be a lower price. (3 stars)</p> <p>3. The lead works great on some tomtoms with the larger USB connector but the adapter supplied doesn't work on the smaller USB connector. Have tried another connector, still no luck, so lead stuck in drawer now!!!! (1 star)</p> <p>4. this item is CREAM & black (NOT white & black)we ordered this online & it was indeed very comfortable & well made. However, it was cream & black in colour, so we returned it and had a look at another in the store that was exactly the same. The manager said he would feed the colour problem back to head office. We will look elsewhere for a black & white beanbag !so - in summary - if you are after a CREAM and black football bean bag then you would be very happy with this product. (2 stars)</p>	<p>1. I suppose it's true that you get what you pay for. It's light and compact, no problem to set up, but the sound quality is very poor. I call it Tin Lizzie. (2 stars)</p> <p>2. Disappointed to say the least. The beads were not aligned and the bar at the top is just a strip of cheap wood that is bent. (1 star)</p> <p>3. Both the store manager and I tried to fit the case, on the date of purchase (it was the appropriate design for my iPod), but the two halves did not fit together. The case was flimsy too, so I don't know how much protection it would have afforded my device anyway. I received my money back there and then. (1 star)</p> <p>4. My son used these twice then they stopped working properly. (1 star)</p>

Table A.6: LDA topics extracted from the Q&A corpus.

Topic	Highest Probability Words
Dimensions	height width depth dimensions length size
Guarantee/warranty	buy bought guarantee product warranty year
Compatibility (computers)	ipod compatible laptop windows work download
Dimensions (furniture) 1	fit sofa flat size item dimensions
Attributes (kitchen appliance)	oven grill light switch time microwave
Dimensions (furniture) 2	door side drawers left doors open
Attributes (furniture)	weight table chair chairs back seat
Installation queries (kitchen)	cooker gas included installation electric include
Compatibility (phone)	phone card sim work memory phones
Compatibility (home appliance)	box work connect record freeview internet
Attributes (lamps)	glass light pole lid lamp plastic
Attributes (power socket)	cable usb plug battery socket power
Attributes (entertainment system)	work player dvd play remote samsung
Dimensions (bed)	bed mattress size fit base double
Queries (home appliance)	fridge freezer free dryer long wash
Dimension (furniture) 3	wall unit shelves shelf top fit
Instructions (camera/printer)	camera printer print clock ink set
Description discrepancies	product description confirm question correct states
Color/finish (furniture)	colour made white black wood match
Attributes (home appliance)	machine water washing make hot filter

Table A.7: LDA topics extracted from the review corpus.

Topic	Highest Probability Words
Quality (vacuum)	easy great good cleaner clean product
Quality (electronics)	sound good great clock quality set
Quality (phone/camera)	phone easy set good camera features
Quality (home appliance)	kettle water good iron machine toaster
Quality (bedclothes)	bed duvet comfortable warm mattress pillows
Value for money 1	good job easy product money price
Product instructions	put instructions wall screws holes fit
Product returns	store item product service delivery back
Replacement	bought years printer replace buy good
Gifts	bought great room son loves year
Quality (clothes line)	put clothes easy good cover sturdy
Dimensions (storage unit)	storage easy small space put unit
Value for money 2	product money great good recommend excellent
Value for money (negative)	it's bit reviews price cheap buy
Quality (garden tools)	good job cut light easy small mower
Discounts	price good quality great bargain sale
Quality (lighting)	light colour nice lovely lamp room
Assembly instructions	easy good put table assemble money
Quality (kitchen appliance)	easy clean great cooker microwave food
Quality (poor)	quality bin back poor plastic cheap

Table A.8: Flexible definition of holdout: The impact of Q&A on ratings using low pre-treatment ratings (≤ 4) as a proxy for fit uncertainty.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1 st stage)	IV	IV (bins)
POST \times Low Rating	0.243*** (0.022)	0.183*** (0.027)		0.204*** (0.036)	
POST \times Hold-out Rating			-0.918*** (0.025)		
POST	-0.045*** (0.009)	-0.022** (0.011)	4.121*** (0.110)	-0.025** (0.011)	
POST \times Rating $\in [2,3]$					0.557*** (0.107)
POST \times Rating $\in (3,4]$					0.128*** (0.031)
POST \times Rating $\in (4,5]$					-0.018* (0.011)
Review Rank	0.0002*** (0.0001)	0.0001* (0.0001)	-0.0001* (0.00005)	0.0001* (0.0001)	0.0001* (0.0001)
Product FE	Yes	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes	Yes
F Statistic			455.98		
Observations	345,168	225,182	225,182	225,182	225,182

Note:

*p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table A.9: Flexible definition of holdout: The impact of Q&A on ratings using high pre-treatment rating variance (≥ 1) as a proxy for fit uncertainty.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1 st stage)	IV
POST \times High Variance	0.157*** (0.015)	0.124*** (0.017)		0.153*** (0.025)
POST \times Hold-out Variance			0.611*** (0.012)	
POST	-0.060*** (0.010)	-0.042*** (0.011)	-0.155*** (0.015)	-0.053*** (0.013)
Review Rank	0.0001*** (0.00004)	0.0001* (0.0001)	0.00001 (0.0001)	0.0001 (0.0001)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			823.99	
Observations	345,168	225,182	225,182	225,182

Note:

*p<0.1; **p<0.05; ***p<0.01

Standard errors clustered at the product level.

Table A.10: Flexible definition of holdout: The impact of Q&A on ratings using the pre-treatment fraction of review mentioning fit issues as a proxy for fit uncertainty.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1 st stage)	IV
POST × Fit	1.933*** (0.150)	1.940*** (0.264)		1.722*** (0.348)
POST × Hold-out Fit			1.440*** (0.041)	
POST	-0.039*** (0.010)	-0.029** (0.012)	-0.007*** (0.001)	-0.025** (0.012)
Review Rank	0.0002*** (0.00005)	0.0002** (0.0001)	-0.00000* (0.00000)	0.0001** (0.0001)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			418.86	
Observations	345,168	225,182	225,182	225,182

Note:

*p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table A.11: LDA topics extracted from the pre-Q&A reviews.

Topic	Highest Probability Words
Furniture 1	table seat present glad black
Value for money 1	excellent money value good product
Quality 1	set feature delight read pretty
Phone	recommend sound phone work highly
Value for money 2	good great easy look price
Replacement	old year bought replace purchase
Storage	small easy fit space storage
Furniture 2	bed comfort chair bought mattress
Accessories	design star rang push piece
Returns	return better connect review work
Usage	easy use said work simple
Holidays	iron lid bag bin christmas
Quality 2	light look love colour nice
Furniture 2	price love table great bought
Kitchen appliance	use heat water cook kettle
Household appliance	use clean floor cleaner vacuum
Instructions	instruct wall screw bit drill
Outdoor equipment	machine product price garden
Clocks	clock keep cheap time real
Furniture 3	cover door easy plenty heavy

A.2 Mathematical Appendix

In this appendix we present a stylized model that captures the essence of how the presence of informative Q&A affects consumer decision making and product ratings in settings with consumer fit uncertainty.

Consumer Side. We begin by modeling the consumer side. Understanding how the presence of Q&A affects consumer behavior is essential in order to understand the impact of Q&A on average product ratings.

A focal consumer contemplates whether to purchase a product. The consumer possesses perfect knowledge about every attribute of the product, except one. The unknown attribute can take one of two values, a “good” value resulting in positive product utility g and a “bad” value resulting in negative utility $-b$ ($g, b \geq 0$). Both utilities g, b include the disutility of price. All utilities, as well as the definition of what is “good” and “bad”, might differ from one consumer to the next. Therefore, our model is general enough to encompass both quality and fit-related attributes. In the latter case, “good” and “bad” have subjective interpretations. For example, “good (bad)” might mean “the dimensions of this product fit (do not fit) through my apartment’s door” or “the lens is (is not) compatible with my camera.”

Let us denote by α the prior probability that the unknown attribute will take the “bad” value; α is thus the probability of “bad fit” or fit mismatch. In the absence of any additional information, the consumer’s expected utility is $g - \alpha(g + b)$. The consumer purchases if and only if $g - \alpha(g + b) > 0$ or, equivalently, if $\alpha < g/(g + b)$. If the consumer purchases, with probability α she experiences negative post-purchase utility, i.e. *regrets* the purchase. We assume that if the consumer experiences positive (negative) post-purchase utility she posts a positive (negative) product review. Assuming that a positive review is equivalent to a rating of “1” and a negative review

equivalent to a rating of “0”, the average product rating is equal to one minus the average probability of post-purchase regret among its purchasers.¹

Let us now assume that the consumer asks a question about the value of the unknown attribute (by posting a question at a Q&A forum) and receives back an answer. The answer can be positive, meaning “the attribute is good” or negative, meaning, “the attribute is bad”. We assume that the answer is correct with probability $p \geq \frac{1}{2}$. Thus, p denotes the quality of information. Denote by π_+, π_- the posterior probabilities that the unknown attribute has the “bad” value given positive (+) or negative (-) answers respectively. According to standard Bayesian inference, it is:

$$\pi_+ = \frac{Pr[+|b]Pr[b]}{Pr[+]} = \frac{(1-p)\alpha}{(1-p)\alpha + p(1-\alpha)} \quad \pi_- = \frac{Pr[-|b]Pr[b]}{Pr[-]} = \frac{p\alpha}{p\alpha + (1-p)(1-\alpha)}$$

It is easy to show that:

- π_+ is monotonically decreasing with p and ranges from α (for $p = \frac{1}{2}$) to 0 (for $p = 1$)
- π_- is monotonically increasing with p and ranges from α (for $p = \frac{1}{2}$) to 1 (for $p = 1$)
- $\pi_+ \leq \alpha \leq \pi_-$ for all $p \geq \frac{1}{2}$

The consumer’s expected utility from purchase, given answer $s \in \{+, -\}$, is equal to:

$$u_s = (1 - \pi_s)g + (\pi_s)(-b) = g - \pi_s(g + b)$$

¹The model can be extended to a multi-valued rating scale $1, 2, \dots, n$ by defining a correspondence between post-purchase utilities u_1, u_2, \dots, u_{n-1} , where $u_i < u_{i+1}$, such that consumers post rating i if they experience post-purchase utility $u_{i-1} < u \leq u_i$ plus the obvious corner cases. The precise thresholds u_i may differ among consumers. Such a mapping retains the key properties that drive our stylized model, i.e. average ratings are positively related to average post-purchase utility and negatively related to the probability of fit mismatch among purchasers.

The consumer purchases if and only if $u_s > 0$. There are two cases:

Case I: (Optimistic consumers) $g - \alpha(g + b) > 0$, or equivalently $\alpha < g/(g + b)$. In this case the consumer would always purchase the product on the basis of her prior beliefs. If we add the option of asking questions, the consumer purchases the product if either: 1) she receives a positive answer of any informativeness, or 2) she receives a negative answer of low informativeness, such that her posterior beliefs remain close to the prior. However, she does not purchase the product if she receives a negative answer whose informativeness p is sufficiently high, such that $\pi_- \geq g/(g + b)$. The consumer's probability of regret conditional on purchase, is equal to α if there is no Q&A or if there is Q&A, as long as the answer's informativeness remains relatively low. The probability of regret, conditional on purchase, *decreases* to π_+ (recall that $\pi_+ \leq \alpha$) as soon as the answer's informativeness p crosses the threshold above which the consumer buys only if she receives a positive answer; π_+ is a declining function of p and converges to zero as p tends towards 1, i.e. as answers to questions become perfectly reliable. In that limiting case, consumers who choose to purchase in the presence of Q&A never experience fit mismatch and post only positive ratings.

Case II: (Pessimistic consumers) $g - \alpha(g + b) \leq 0$, or equivalently $\alpha \geq g/(g + b)$. In this case, the consumer would not purchase the product on the basis of her prior beliefs. If we add the option of asking questions, the consumer only purchases the product if she receives a positive answer whose informativeness p is sufficiently high, such that $\pi_+ \leq g/(g + b)$. If the consumer purchases, the probability of post-purchase regret is π_+ ; as above, the probability of regret goes to zero in the limiting case of perfectly reliable answers.

The conditions that determine whether Case I or Case II applies depend on both α and the ratio $g/(g + b)$. Case I applies when either α is small or $g/(g + b)$ is large.

Case II applies when α is large or $g/(g+b)$ is small. The ratio $g/(g+b)$ captures the relationship between the utility of a match g and the disutility of a mismatch b . In settings where the consequences of a mismatch are not very severe (e.g. when product prices are low and/or products are easy to return), b is likely to be small and $g/(g+b)$ large. Conversely, in settings where the consequences of a mismatch are more severe (e.g. high prices, difficult to return products, bad fit causes damage to property or health) b is likely to be large and $g/(g+b)$ small.

In summary:

1. The presence of Q&A affects consumer decision making only if answers are sufficiently informative (i.e. if p is sufficiently high).
2. The ability to ask a question and receive a sufficiently informative answer about an unknown fit-related product feature has the following effect on consumer decision making:
 - (a) In settings where the prior probability of bad fit is low or the consequences of fit mismatch not severe (Case I), it discourages consumers who would otherwise be making a mistake from purchasing the product if the answer indicates that the product may not be a good fit for them.
 - (b) In settings where the prior probability of bad fit is high or the consequences of fit mismatch severe (Case II), it encourages consumers who would otherwise be reluctant to purchase the product if the answer indicates that the product may be a good fit for them.

Observe that, in our model, average consumer utility and average ratings are linear transformations of one another. Specifically, average consumer utility $u = g - \pi(g+b)$

corresponds to average product rating $r = 1 - \pi$, which gives:

$$r = \frac{1}{g+b}u + \frac{b}{g+b}$$

Product Side. We now turn our attention to the product side. For any given product, we assume that there are multiple prospective consumers. We, further, assume that a fraction $1 - \epsilon$ of consumers (we will call them the “informed” consumers) have perfect information about the product and purchase it, knowing that it serves their needs. These consumers always post positive reviews. The remaining consumers behave like the focal consumer we analyzed above. We will call those consumers the “uninformed” consumers.

Each uninformed consumer may care about different unknown attributes and may have different notions of what constitutes “good” and “bad” states of those attributes. On aggregate, we assume that the focal product is a bad *ex-post* fit for a fraction ω of uninformed consumers and a good fit for the rest. However, uninformed consumers do not have precise fit information *ex-ante* and, as discussed above, make decisions assuming a prior probability of bad fit equal to α .

Note that there is no inconsistency in assuming different values for α and ω . Whereas the value of α reflects the distribution of product attributes on the market, ω reflects the distribution of consumer tastes for those attributes. For example, consider the case of portable hard drives that can be compatible with PC only or Mac only. Assume that 70% of portable hard drives on the market are compatible with PC only and 30% compatible with Mac only. If the drive’s compatibility is the unknown attribute, a PC user would be justified in assuming $\alpha = 0.3$ whereas a Mac user would be justified in assuming $\alpha = 0.7$. Assume, now, that 90% of consumers are PC users and 10% are Mac users. A PC-compatible hard drive is, thus, a bad fit for a fraction $\omega = 0.1$ of consumers, whereas a Mac-compatible hard drive is a bad fit

for a fraction $\omega = 0.9$.

Under the above assumptions, and assuming that Q&A is informative enough (i.e. that p is sufficiently high) to affect consumer behavior:

1. If $\alpha < g/(g+b)$, such that Case I applies:
 - (a) Without Q&A, all uninformed consumers purchase. A fraction ω will experience bad fit and will post negative reviews. The average ratings of uninformed consumers will then be $1 - \omega$ and the average ratings of all consumers (informed plus uninformed) $1 - \epsilon + \epsilon(1 - \omega)$.
 - (b) With Q&A, with probability p a fraction ω of uninformed consumers (the fraction for whom the product is a bad fit) will receive a (correct) negative answer and will not purchase. With probability $1 - p$ that same fraction will receive a (wrong) positive answer and will purchase, resulting in bad fit and negative reviews. With probability p the remaining fraction (the fraction for whom the product is a good fit) will receive a (correct) positive answer and will purchase, resulting in positive reviews. With probability $1 - p$ that same fraction will receive a (wrong) negative answer and will not purchase. The average ratings of uninformed consumers will thus be $p(1 - \omega)/[(1 - p)\omega + p(1 - \omega)]$ and the average ratings of all consumers $(1 - \epsilon + \epsilon p(1 - \omega))/[1 - \epsilon + \epsilon((1 - p)\omega + p(1 - \omega))]$.
2. If $\alpha \geq g/(g+b)$, such that Case II applies.
 - (a) Without Q&A, no uninformed consumers purchase. Informed consumers always post positive ratings, therefore, the average product ratings will be 1.
 - (b) With Q&A, the effect will be identical to case 1(b) above.

Without loss of generality, we assume that a fraction γ of uninformed consumers have prior beliefs α such that Case I applies and the rest have prior beliefs such that Case II applies. Then, combining Cases 1 and 2 above, we conclude that:

- Without Q&A (Cases 1(a) and 2(a)), the average rating of the product will be $(1 - \epsilon + \epsilon\gamma(1 - \omega))/(1 - \epsilon + \epsilon\gamma)$. Using elementary comparative statics we can show that this is a decreasing function of ϵ , ω and γ , that is, ratings are lower when:
 1. there are many uninformed consumers (high ϵ), that is, the product exhibits a higher fit uncertainty, and
 2. the product is a bad fit for a large fraction of consumers (high ω), that is, the product caters to niche tastes that do not coincide with the mainstream², and
 3. many uninformed consumers are optimistic about the probability of a good fit and choose to purchase in the presence of fit uncertainty (high γ); as previously discussed, this happens when most products of this category are a good fit for most consumers (such that most consumers have a low α) and/or when the impact of bad fit is not very severe relative to the utility of a good fit.

- An interesting nuance of this result is that *high* average ratings may indicate one or more of the following conditions:
 1. there exist few uninformed consumers (low ϵ)
 2. the product is a good fit for a lot of consumers (low ω)

²Assuming that most computer users are PC users, an external hard drive that is only compatible with Apple computers would be an example of such a niche product.

3. most uninformed consumers are pessimistic and choose to not purchase; as previously discussed, this happens when the probability of fit mismatch is high for this product category (high α) and/or the impact of bad fit is severe relative to the utility of a good fit
- With sufficiently informative Q&A (Cases 1(b) and 2(b)), γ does not matter and the average rating of the product will be $(1 - \epsilon + \epsilon p(1 - \omega)) / [1 - \epsilon + \epsilon((1 - p)\omega + p(1 - \omega))]$. Using elementary comparative statics we can show that this is an increasing function of p and a decreasing function of ϵ and ω . Average ratings are higher, the higher the informativeness of Q&A and the lower the fraction of 1) uninformed consumers and 2) consumers for which the product is not a good fit.

The crispest intuitions are obtained when Q&A answers are always correct ($p = 1$). Average ratings with Q&A are then equal to 1 for all ω , since consumers become perfectly informed and purchase if and only if the product is a good fit for them. The ratings increase due to Q&A is then simply $1 - (\text{Ratings without Q\&A}) = \epsilon\gamma\omega / (1 - \epsilon + \epsilon\gamma)$.³ The latter is an increasing function of ϵ , ω and γ . The ratings increase is also inversely proportional to the ratings without Q&A, i.e. the lower these ratings, the higher the increase.

To reiterate, the positive impact of Q&A on average ratings of individual products is highest for products that

1. exhibit fit uncertainty for many consumers, and
2. are not a good fit for many consumers, and
3. belong to product categories where many uninformed consumers are optimistic about the probability of a good fit and choose to purchase in the presence of fit

³This expression roughly corresponds to m_j in our empirical model.

uncertainty; as previously discussed, this happens when most products of this category are a good fit for most consumers and/or when the impact of bad fit is not very severe relative to the utility of a good fit.

When $p < 1$ the ratings increase due to Q&A is generally lower and may even become negative in settings where γ is close to 0. Such settings are characterized by pessimistic uninformed consumers who have very unfavorable priors about product fit or are faced with severe consequences of fit mismatch (i.e. fall in Case II of the consumer model). In the absence of Q&A, pessimistic uninformed consumers do not purchase; only informed consumers purchase and post positive ratings. The presence of Q&A may convince uninformed consumers to purchase if they receive a positive answer. However, because this answer may be wrong with some probability, some uninformed purchasers will experience bad fit and will post negative ratings, thus lowering the (previously perfect) average. In the paper we assume that p is sufficiently close to one (and/or γ sufficiently high) for such effects to *not* occur. Our empirical analyses are consistent with these assumptions; we find no statistically significant evidence of average ratings *declining* after questions are answered.

Appendix B

Reference Price Effects

B.1 Doubly Robust Estimation of Heterogeneous Treatment Effects

The Doubly Robust approach flexibly applies machine learning models to create individual level estimates of treatment effects. Resultantly, these estimates can then be projected onto the space of covariates for which we wish to model heterogeneity (while marginalising over all other covariates). All analysis was conducted using the EconML package developed by Microsoft Research.¹

We assume that:

$$\begin{aligned} Y^{(t)} &= g_t(X, W) + \epsilon_t, & \mathbb{E}[\epsilon | X, W] &= 0 \\ Y^{(t)} &\perp T | X, W \end{aligned} \tag{B.1}$$

Hence, modifying Equation B.1 we have:

$$\theta(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X] = \mathbb{E}[g_1(X, W) - g_0(X, W)|X] \tag{B.2}$$

One way to estimate $\theta(x)$ is the Direct Method (DM) approach, where we compute $Y_i(DM)$ as $Y_{i,t}^{DM} = g_t(X_i, W_i) - g_0(X_i, W_i)$. The task then amounts to estimating $g(X, W)$ using machine learning methods and then evaluating $\theta(X)$ by regressing $Y_{i,t}^{DM}$ on X . The main problem with this approach is that it is heavily dependent on the model-based extrapolation that is implicitly done via the model that is fitted in

¹<https://econml.azurewebsites.net/>

the regression.

An alternative approach that does not suffer from the aforementioned problems is the Inverse Propensity Score (IPS) approach. This method starts from the realization that, due to the unconfoundedness assumption, we can create an unbiased estimate of every potential outcome by re-weighting each sample by the inverse probability of that sample receiving the treatment we observed (i.e. up-weighting samples that have “surprising” treatment assignments).

$$Y_{i,t}^{IPS} = \frac{Y_i 1\{T_i = t\}}{\Pr[T_i = t | X_i, W_i]} = \frac{Y_i 1\{T_i = t\}}{p_t(X_i, W_i)} \quad (\text{B.3})$$

Then it holds that

$$\begin{aligned} \mathbb{E}[Y_{i,t}^{IPS} | X, W] &= \mathbb{E}\left[\frac{Y_i 1\{T_i = t\}}{p_t(X_i, W_i)} \mid X_i, W_i\right] = \mathbb{E}\left[\frac{Y_i^{(t)} 1\{T_i = t\}}{p_t(X_i, W_i)} \mid X_i, W_i\right] \\ &= \mathbb{E}\left[\frac{Y_i^{(t)} \mathbb{E}[1\{T_i = t\} \mid X_i, W_i]}{p_t(X_i, W_i)} \mid X_i, W_i\right] = \mathbb{E}[Y_i^{(t)} \mid X_i, W_i] \end{aligned} \quad (\text{B.4})$$

Thus we can estimate a $\theta(X)$ by regressing $Y_{i,t}^{IPS} - Y_{i,0}^{IPS}$ on X . The drawback with this approach is that it has high variance, since we are dividing the observation by a relatively small probability (especially if in some regions of X, W some treatments are unlikely).

The Doubly Robust approach avoids the above drawbacks by combining the two methods. In particular, it fits a direct regression model, but then debiases that model by applying an Inverse Propensity approach to the residual of that model. It constructs the following estimates of the potential outcomes:

$$Y_{i,t}^{DR} = g_t(X_i, W_i) + \frac{Y_i - g_t(X_i, W_i)}{p_t(X_i, W_i)} \cdot 1\{T_i = t\} \quad (\text{B.5})$$

Then we can learn $\theta(x)$ by regressing $Y_{i,t}^{DR} - Y_{i,0}^{DR}$ on X_i .

This yields the overall algorithm:

1. Learn a ‘regression model’ $g^t(X, W)$, by running a regression of Y_i on T_i, X_i, W_i
2. Learn a ‘propensity model’ $p^t(X, W)$, by running a classification to predict T_i from X_i, W_i
3. Construct the doubly robust random variables as described above and regress them on suitable X_i s to explore heterogeneity.

We apply the DR algorithm by using a Lasso as the first stage model, and a logistic regression as the propensity model. In both case, all available pre-treatment covariates enter as predictors. Estimation is performed in a cross-fitting way (i.e, the sample used to fit the predictive models is decoupled from the samples used in the final regression, as suggested by e.g Chernozhukov et al. (2018); Athey and Wager (2021))². We also perform 3-fold cross validation for parameter tuning. After obtaining individual level treatment effect estimates in line with Equation B.5 for each i , we can then project these onto different sets of covariates to provide estimates of CATEs.

²Cross-fitting means that $g_t(X_i, W_i)$ is estimated without using individual i 's own data in the training process. We can split data randomly into n folds - then $g_t(X_i, W_i)$ for individuals in a given fold is trained only using data from the other $n - 1$ folds. This reduces over-fitting and improves efficiency. We use $n = 3$ in our estimation.

B.2 CATE estimates: doubly robust

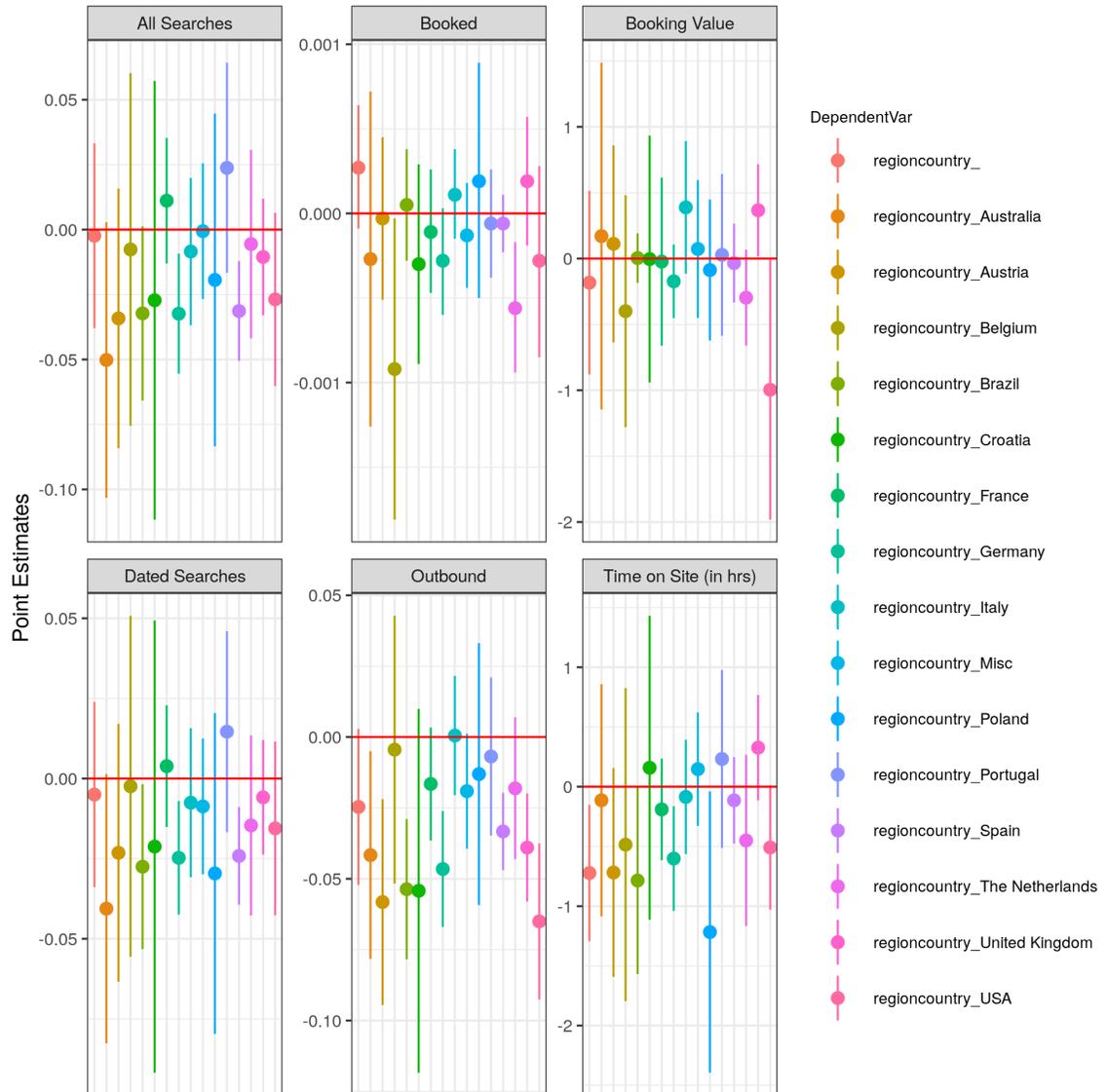


Figure B.1: Region Country

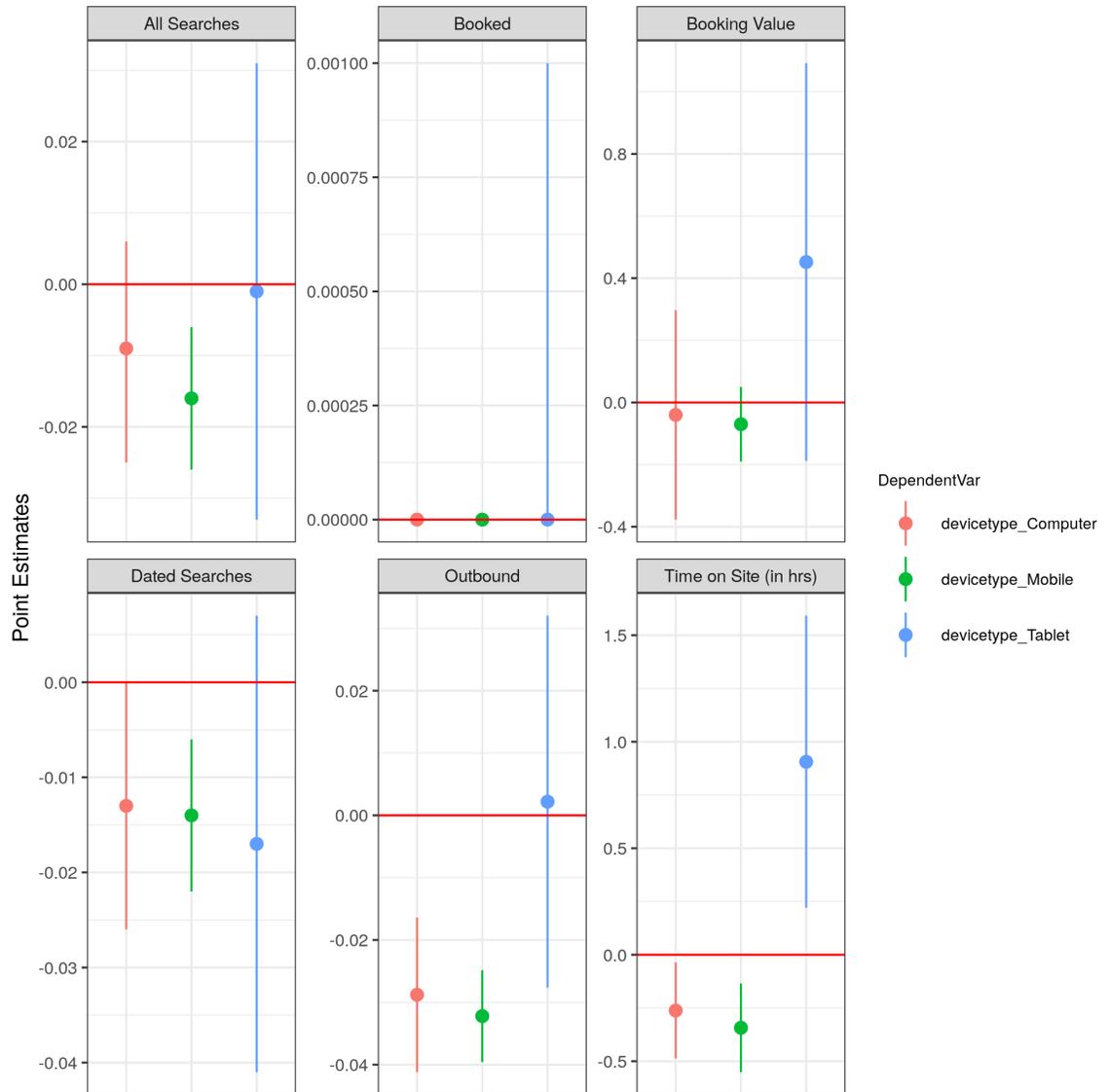


Figure B·2: Device

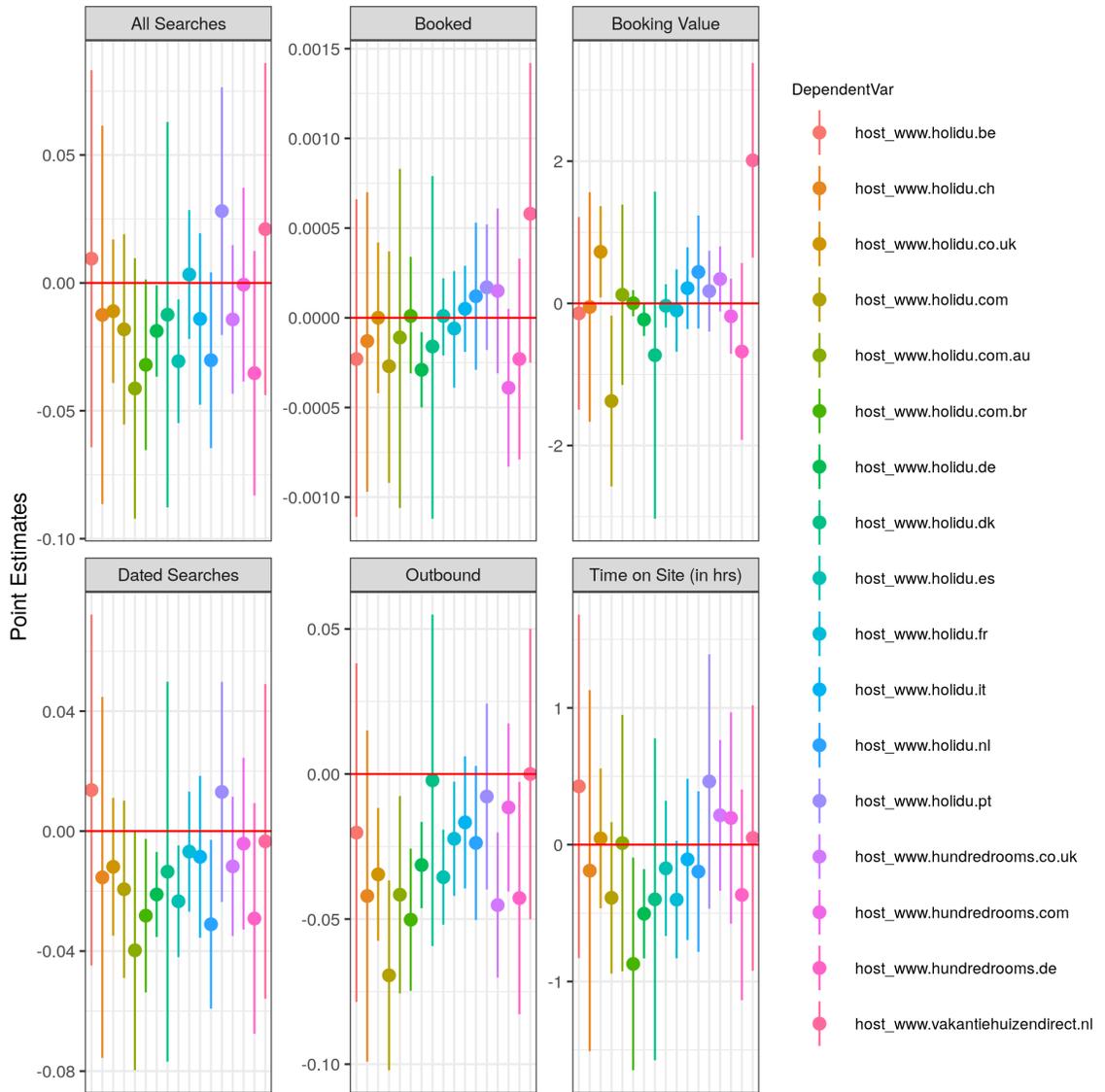


Figure B-3: Host

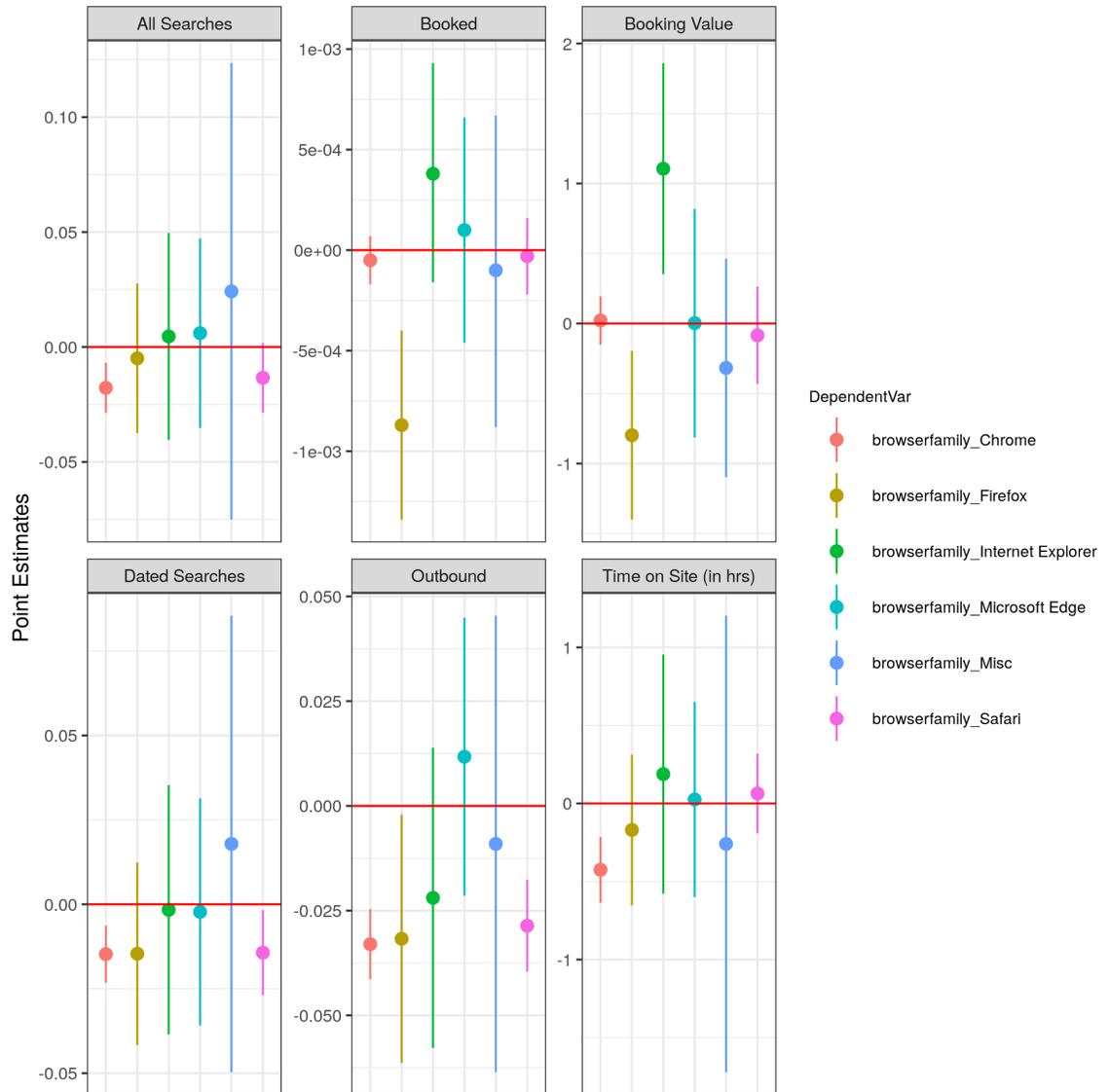


Figure B-4: Browser

B.3 Some correlational evidence from Airbnb

As a follow up to our experimental results, here we provide some results using floor prices on Airbnb.com. In particular, this data gives us an opportunity to explore settings where prices are presented in calendar form and not just revealed when dates are entered. Interestingly, we find that floor prices tend to correlate positively with occupancy rates and booking probability in this setting. While we do not have enough evidence to make a strong causal claim with this data, we provide it as a puzzle that can hopefully stimulate researchers to investigate it further in the future.

“From” prices on Airbnb.com are calculated based on the minimum price that appears on the calendar of a listing over the next 30 days. We examine how this minimum price affects bookings. We collect listing-level calendar data on Airbnb (scraped weekly for 2 months in 2015), with each scrape being associated with a minimum price. We have detailed information on roughly 23,000 listings in the greater Boston and New York area, and static as well as dynamic characteristics (location, size and number of rooms being offered, accommodation type, prices, and ratings). In total, we have 91,268 unique listing-scrape combinations. The identifying variation comes from scrape-to-scrape changes in the floor price for a given listing. Although we cannot observe consumer-level variables for this analysis, this setting allows us to look at the effect of floor prices on the occupancy rate of hotels, since we have access to the entire calendar. As mentioned before, we also know the price associated with every date on the calendar (price information appears next to the calendar date, and was also visible to consumers at the time the data was collected). So, this setting allows us to model the difference between focal and floor prices for every date, and examine how this gap affects bookings. Our observational results indicate that, a larger floor price, as well as a smaller relative difference between the displayed From price and the actual price paid, leads to both greater booking probability and higher occupancy

rates, while controlling for focal prices and various time-invariant fixed effects, as well as different functional forms. To further alleviate concerns of endogeneity, we estimate an additional specification that compares listings with a positive change in From price over the 2 month observation period and those with no change. The listings that see a change of greater than \$5 are labeled as “treatment” listings, and matched with control listings based on a set of listing level characteristics (we use robust Lasso (Chernozhukov et al., 2016) to control for selection on observables). We find that treatment listings have a higher occupancy rate relative to control.

We start with a descriptive plot that examines mean booking probability as a function of the wedge (difference between focal price and minimum (floor) price). First, we bin the wedge in bins of 10. Then, we calculate the percentage of booked nights within each bin. Finally, we create a plot of this percentage against the bins. We see a declining trend as the difference is larger (Figure B·5).

B.3.1 Within listings

First, our goal is to estimate the impact of the floor price, over and above the focal price and other listing and time level characteristics. To do so, we make use of three key independent variables :

1. the “Starting from..’ price associated with a given listing-calendar date ($Price_{From}$)
2. the absolute difference (Diff) between the focal and the “Starting from..” price and,
3. the relative difference ($Frac.Diff = \frac{Diff}{Price}$) between the focal and the “Starting from..” price.

In our first set of regressions, our dependent variable of interest in an binary

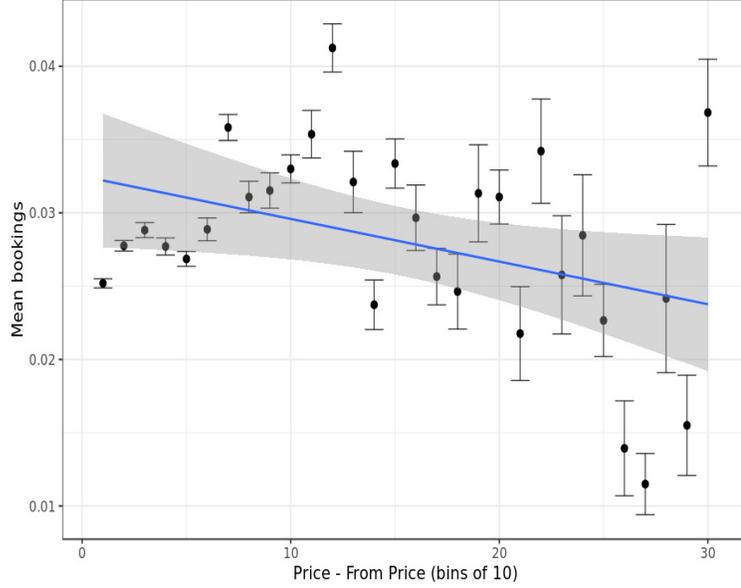


Figure B-5: Effect of the wedge difference in Prices and floor prices on booking probability. First, differences are binned in groups of 10. Then, for each bin, the average booking probability is computed. We see that the probability decreases as the difference increases. The bars indicate 95% confidence intervals, and the fitted line and shaded region plots out smoothed conditional means.

indicator for whether a listing i has a booked night j at time t . The unit of analysis is thus at the listing-scrape-calendar date level. The equations estimated are:

$$\begin{aligned}
 \text{Booking}_{ijt} &= \beta_1 \text{Price}_{Fromit} + \beta_2 \text{Price}_{ijt} + \alpha_i + \delta_t + \epsilon_{ijt} \\
 \text{Booking}_{ijt} &= \beta_1 \text{Diff}_{ijt} + \beta_2 \text{Price}_{ijt} + \alpha_i + \delta_t + \epsilon_{ijt} \\
 \text{Booking}_{ijt} &= \beta_1 \text{Frac.Diff}_{ijt} + \beta_2 \text{Price}_{ijt} + \alpha_i + \delta_t + \epsilon_{ijt}
 \end{aligned}
 \tag{B.6}$$

We include fixed effects for listing, month of scrape and month of calendar, to account for time invariant characteristics and transient time shocks respectively. Standard errors are clustered at the listing level. The results are reported in Table B.1. We see that larger From price, as well as a smaller relative difference between the displayed From price and the actual price paid, leads to greater booking probability. Column (1) uses Price and Price_{From} as independent variables; column (2) looks at

their difference relative to Price and column (3) looks at their absolute difference.

Table B.1: A larger From price, as well as a smaller relative difference between the displayed From price and the actual price paid, leads to greater booking probability. The unit of analysis is a listing-scrape-calendar day. Column (1) uses Price and Price_{From} as independent variables; column (2) looks at their difference relative to Price and column (3) looks at their absolute difference

	<i>Dependent variable:</i>		
	Booked (binary)		
	(1)	(2)	(3)
Price _{From}	0.00003 (0.00001)** p = 0.029		
$\frac{\text{Price} - \text{Price}_{\text{From}}}{\text{Price}}$		-0.014 (0.003)** p = 0.00003	
Price - Price _{From}			-0.00003 (0.00001)** p = 0.029
Price	-0.00001 (0.00001)* p = 0.097	0.00002 (0.00001)** p = 0.036	0.00001 (0.00001) p = 0.252
Listing Fixed Effect	Yes	Yes	Yes
Travel month Fixed Effect	Yes	Yes	Yes
Scrape month Fixed Effect	Yes	Yes	Yes
Observations	18,967,778	18,967,778	18,967,778
Adjusted R ²	0.152	0.152	0.152

Note:

*p<0.1; **p<0.05; ***p<0.01

Next, instead of looking at booking probabilities, we look at the occupancy rate (defined as the fraction of booked nights out of all available nights). The unit of analysis here is aggregated at the listing-scrape level. Hence, for every listing i and every scrape instance t , we take averages of the variables of interest over the entire calendar. Hence, for every listing-scrape, we look at average of Price, Diff and Frac.Diff

(Price_{From} is already at the listing-scrape level and can thus enter as is):

$$\begin{aligned} \text{OccupancyRate}_{it} &= \beta_1 \text{Price}_{From_{it}} + \beta_2 \overline{\text{Price}_{it}} + \alpha_i + \delta_t + \epsilon_{it} \\ \text{OccupancyRate}_{it} &= \beta_1 \overline{\text{Diff}_{it}} + \beta_2 \overline{\text{Price}_{it}} + \alpha_i + \delta_t + \epsilon_{it} \\ \text{OccupancyRate}_{it} &= \beta_1 \overline{\text{Frac.Diff}_{it}} + \beta_2 \overline{\text{Price}_{it}} + \alpha_i + \delta_t + \epsilon_{it} \end{aligned} \tag{B.7}$$

The results are reported in Table B.2. Again, we see that a larger From price, as well as a smaller relative difference between the displayed From price and the actual price paid, leads to higher occupancy rate. Column (1) uses $\overline{\text{Price}}$ and Price_{From} as independent variables; column (2) looks at their mean absolute difference (by scrape), and column (3) looks at their mean difference relative to Price (by scrape).

A possible concern with the above specifications might be the high correlation between prices and “Starting from...” prices. We follow Jindal (2018) in designing our key specifications - however, as with any specification that control for focal prices while also modeling the gap between price and reference prices, multicollinearity might be an issue. However, it is to be noted that this would inflate our standard errors, and thus make it harder to find significant effects. Hence, finding significance in this setting is eventually a more stringent test of our hypotheses.

Further, we are concerned about the endogeneity of prices. Given we are modeling demand as a function of prices, there is a risk of simultaneity/omitted variables that biases the price coefficient. This might partially explain the lack of significance of the price coefficient. The key omitted variable in our setting is likely to be some measure of absolute quality - quality would be positively correlated with price, and with demand, which would lead to our observed effects. Unfortunately, we cannot control for quality using a rating metric, due to the short time series of scrapes. Ratings on Airbnb are inflated overall, and there is hardly any variation in star ratings over a 2 month period.

Table B.2: A larger From price, as well as a smaller relative difference between the displayed From price and the actual price paid, leads to higher occupancy rate (calculated as the fraction of total available calendar days that are booked).

	<i>Dependent variable:</i>		
	Occupancy Rate		
	(1)	(2)	(3)
$\overline{\text{Price}}$	-0.0002 (0.00004)*** p = 0.00000	0.00000 (0.00002) p = 0.887	-0.00000 (0.00002) p = 0.823
$\text{Price}_{\text{From}}$	0.0002 (0.00003)*** p = 0.00000		
$\overline{\text{Price} - \text{Price}_{\text{From}}}$		-0.0002 (0.00003)*** p = 0.00000	
$\frac{\overline{\text{Price} - \text{Price}_{\text{From}}}}{\overline{\text{Price}}}$			-0.068 (0.011)*** p = 0.000
Listing Fixed Effect	Yes	Yes	Yes
Scrape month Fixed Effect	Yes	Yes	Yes
Observations	89,491	89,491	89,491
Adjusted R ²	0.147	0.147	0.148

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

Finally, another concern is related to the effect sizes. The means of our two dependent variables (booking probability and occupancy rate) as 0.02 and 0.036 respectively. Hence, our effect sizes are < 1%. However, it is to be noted that, given we make use of observational variation within listings, the changes are fairly small (the mean value of standard deviation for “Starting from...” prices within listing is \$5). This suggests that an average change of as little as \$5 in the From price can affect booking outcomes, which highlights that the results are economically significant.

B.3.2 Across listings

Given the above limitations, we now report results from an across-listing identification strategy. We pick listings that experience a large change in From prices (\$5 and more) as the “treated” listings. To ensure comparability between treated and control listings, we add covariates for (1) neighborhood, (2) number of beds, (3) bed type, (4) city, (5) whether instant bookable, (6) price for an extra person, (6) number of reviews, (7) number of pictures and (8) person capacity. We could not use star rating as a control because Airbnb only releases ratings once listings get more than 3 reviews (which most listings in our sample do not). However, conditional on the above covariates, the mean price across treatment and control listings are not significantly different. Thus, we can argue that we are capturing underlying listing quality with these variables. The results are reported in Table B.3 - column (1) has no controls and column (2) implements variable selection using RLasso³ (Chernozhukov et al., 2016). Again, we find that treated listings (which experience an increase in From price) have a higher occupancy rate relative to control.

Thus, we find evidence that “From” price levels are positively correlated with booking probability and occupancy rates at the listing level.

³This approach allows us to obtain valid confidence bounds on a causal/target parameter while flexibly controlling for “nuisance” parameters (in our case, listing covariates)

Table B.3: Occupancy rate for “treated” listings with a positive change in min price (≥ 5) vs those with smaller or no change.

	<i>Dependent variable:</i>	
	Occupancy Rate	
	(1)	(2)
Treated	0.005 (0.002) ^{***} p = 0.003	0.003 (0.0015) [*] p = 0.058
Listing level covariates	No	Yes -rlasso
Observations	20,396	20,396
Adjusted R ²	0.0004	0.009
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01	

Bibliography

- Acemoglu, D. and Finkelstein, A. (2008). Input and technology choices in regulated industries: Evidence from the health care sector. *Journal of Political Economy*, 116(5):837–880.
- Agichtein, E., Liu, Y., and Bian, J. (2009). Modeling information-seeker satisfaction in community question answering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(2):10.
- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley & Sons Inc., Publication.
- Allender, W. J., Liaukonyte, J., Nasser, S., and Richards, T. J. (2021). Price fairness and strategic obfuscation. *Marketing Science*, 40(1):122–146.
- Anderson, E. T. and Simester, D. I. (2010). Price stickiness and customer antagonism. *The Quarterly Journal of Economics*, 125(2):729–765.
- Angrist, J. D. and Imbens, G. W. (1995). Average causal response with variable treatment intensity. *NBER Working Paper*, <https://www.nber.org/papers/t0118>, (t0118).
- Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist’s companion*. Princeton university press.
- Archak, N., Ghose, A., and Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8):1485–1509.
- Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.
- Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

- Bearden, W. O., Lichtenstein, D. R., and Teel, J. E. (1984). Comparison price, coupon, and brand effects on consumer reactions to retail newspaper advertisements. *Journal of Retailing*, 60(2):11–34.
- Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., and Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *International Conference on Web and Social Media*: <https://www.icwsm.org/papers/3--Benamara-Cesarano-Picariello-Reforgiato-Subrahmanian.pdf>. Cite-seer.
- Blake, T., Moshary, S., Sweeney, K., and Tadelis, S. (2018). Price salience and product choice. *NBER Working Paper*, <https://www.nber.org/papers/w25186>.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning research*, 3(Jan):993–1022.
- Bondi, T. (2019). Alone, together: Product discovery through consumer ratings. Available at SSRN 3468433.
- Branco, F., Sun, M., and Villas-Boas, J. M. (2015). Too much information? information provision and search costs. *Marketing Science*, 35(4):605–618.
- Burke, R. R., DeSarbo, W. S., Oliver, R. L., and Robertson, T. S. (1988). Deception by implication: An experimental investigation. *Journal of Consumer Research*, 14(4):483–494.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C48. <https://doi.org/10.1111/ectj.12097>.
- Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2017). Generic machine learning inference on heterogenous treatment effects in randomized experiments. arxiv e-prints. *arXiv preprint arXiv:1712.04802*.
- Chernozhukov, V., Hansen, C., and Spindler, M. (2016). High-dimensional metrics in r. *arXiv preprint arXiv:1603.01700*.
- Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American economic review*, 99(4):1145–77.
- Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

- Darke, P. R. and Ritchie, R. J. (2007). The defensive consumer: Advertising deception, defensive processing, and distrust. *Journal of Marketing research*, 44(1):114–127.
- Della Bitta, A. J., Monroe, K. B., and McGinnis, J. M. (1981). Consumer perceptions of comparative price advertisements. *Journal of Marketing Research*, 18(4):416–427.
- Dhanasobhon, S., Chen, P.-Y., Smith, M., and Chen, P.-y. (2007). An analysis of the differential impact of reviews and reviewers at amazon. com. *International Conference in Information Systems 2007 Proceedings*, page 94.
- Dimoka, A., Hong, Y., and Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence. *MIS quarterly*, 36.
- Donald, S. G. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233.
- Ellison, G. (2005). A model of add-on pricing. *The Quarterly Journal of Economics*, 120(2):585–637.
- Ellison, G. and Ellison, S. F. (2018). Search and obfuscation in a technologically changing retail environment: Some thoughts on implications and policy. *Innovation Policy and the Economy*, 18(1):1–25.
- Foster, D. J. and Syrgkanis, V. (2019). Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*.
- Gabaix, X. and Laibson, D. (2006). Shrouded attributes, consumer myopia, and information suppression in competitive markets. *The Quarterly Journal of Economics*, 121(2):505–540.
- Gallino, S. and Moreno, A. (2018). The value of fit information in online retail: Evidence from a randomized field experiment. *Manufacturing & Service Operations Management*, 20(4):767–787.
- Godes, D. and Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473.
- Griliches, Z. and Hausman, J. A. (1986). Errors in variables in panel data. *Journal of econometrics*, 31(1):93–118.
- Gupta, A. (2021). Impacts of performance pay for hospitals: The readmissions reduction program. *American Economic Review*, 111(4):1241–83.
- Hong, Y. and Pavlou, P. A. (2014). Product fit uncertainty in online markets: nature, effects, and antecedents. *Information Systems Research*, 25(2):328–344.

- Huang, J. Z. J. (2018). The thrill of the deal : Quantifying the price of perceived discounts and markups. https://marketing.wharton.upenn.edu/wp-content/uploads/2018/09/09.20.2018-Huang-Jennie-PAPER-ThrilloftheDeal_JennieHuang.pdf.
- Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.
- Jindal, P. (2018). Reference dependence and price negotiations—the role of advertised reference prices. Available at SSRN: <https://ssrn.com/abstract=3189448>.
- Khern-am nuai, W., Ghasemkhani, H., Qiao, D., and Kannan, K. N. (2020). The impact of online q&as on product sales: The case of amazon answer. Available at SSRN 2794149.
- Kopalle, P. K. and Lindsey-Mullikin, J. (2003). The impact of external reference price on consumer price expectations. *Journal of Retailing*, 79(4):225–236.
- Kwark, Y., Chen, J., and Raghunathan, S. (2014). Online product reviews: Implications for retailers and competing manufacturers. *Information systems research*, 25(1):93–110.
- Lappas, T., Dellarocas, C., and Derakhshani, N. (2017). Reputation and contribution in online question-answering communities. Available at SSRN: <https://ssrn.com/abstract=2918913>.
- Lichtenstein, D. R. and Bearden, W. O. (1989). Contextual influences on perceptions of merchant-supplied reference prices. *Journal of Consumer Research*, 16(1):55–66.
- Lichtenstein, D. R., Burton, S., and Karson, E. J. (1991). The effect of semantic cues on consumer perceptions of reference price ads. *Journal of consumer research*, 18(3):380–391.
- Liu, X., Lee, D., and Srinivasan, K. (2019). Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Journal of Marketing Research*, 56(6):918–943.
- Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp.com. *HBR Working Paper 12-016*: <https://www.hbs.edu/faculty/Pages/item.aspx?num=41233>.
- Mamadehussene, S. (2020). The interplay between obfuscation and prominence in price comparison platforms. *Management Science*, 66(10):4359–4919.
- Manchanda, P., Packard, G., and Pattabhiramaiah, A. (2015). Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Marketing Science*, 34(3):367–387.

- Mayhew, G. E. and Winer, R. S. (1992). An empirical analysis of internal and external reference prices using scanner data. *Journal of Consumer Research*, 19(1):62–70.
- Mazumdar, T., Raj, S. P., and Sinha, I. (2005). Reference price research: Review and propositions. *Journal of marketing*, 69(4):84–102.
- McAuley, J. and Yang, A. (2016). Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.
- Monroe, K. B. (1973). Buyers’ subjective perceptions of price. *Journal of marketing research*, pages 70–80.
- Morwitz, V. G., Greenleaf, E. A., and Johnson, E. J. (1998). Divide and prosper: consumers’ reactions to partitioned prices. *Journal of Marketing Research*, 35(4):453–463.
- Narayanan, S. and Nair, H. S. (2013). Estimating causal installed-base effects: A bias-correction approach. *Journal of Marketing Research*, 50(1):70–94.
- Ngwe, D. (2018). Fake discounts drive real revenues in retail. *Harvard Business School*.
- Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, pages 1417–1426.
- Proserpio, D. and Zervas, G. (2017). Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science*, 36(5):645–665.
- Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.
- Ravi, S., Pang, B., Rastogi, V., and Kumar, R. (2014). Great question! question quality in community q&a. In *International Conference of Web and Social Media*.
- Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical Science*, pages 15–32.
- Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.
- Roodman, D. (2009). How to do xtabond2: An introduction to difference and system gmm in stata. *The stata journal*, 9(1):86–136.

- Sahoo, N., Dellarocas, C., and Srinivasan, S. (2018). The impact of online product reviews on product returns. *Information Systems Research*, 29(3):723–738.
- Senecal, S. and Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, 80(2):159–169.
- Sullivan, M. W. (2017). Economic analysis of hotel resort fees. *Economic Issues, Bureau of Economics Federal Trade Commission*.
- Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, 58(4):696–707.
- Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., and Macke, J. H. (2020). sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505.
- Thaler, R. (1985). Mental accounting and consumer choice. *Marketing science*, 4(3):199–214.
- Tirole, J. (1988). *The Theory of Industrial Organization*. MIT Press.
- Urbany, J. E., Bearden, W. O., and Weilbaker, D. C. (1988). The effect of plausible and exaggerated reference prices on consumer perceptions and price search. *Journal of consumer research*, 15(1):95–110.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Winer, R. S. (1986). A reference price model of brand choice for frequently purchased products. *Journal of Consumer Research*, 13(2):250–256.
- Xia, L., Monroe, K. B., and Cox, J. L. (2004). The price is unfair! a conceptual framework of price fairness perceptions. *Journal of Marketing*, 68(4):1–15.
- Zhu, F. and Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74(2):133–148.

CURRICULUM VITAE

