ABSTRACT

Title of Dissertation:	ASSESSING QUALITY IN HIGH-UNCERTAINTY MARKETS: ONLINE REVIEWS OF CREDENCE SERVICES
	Shannon Lantzy, Doctor of Philosophy, 2016
Dissertation directed by:	Associate Professors of Information, Katherine Stewart & Siva Viswanathan, Decision, Operations & Information Technologies

In economics of information theory, credence products are those whose quality is difficult or impossible for consumers to assess, even after they have consumed the product (Darby & Karni, 1973). This dissertation is focused on the content, consumer perception, and power of online reviews for credence services. Economics of information theory has long assumed, without empirical confirmation, that consumers will discount the credibility of claims about credence quality attributes. The same theories predict that because credence services are by definition obscure to the consumer, reviews of credence services are incapable of signaling quality. Our research aims to question these assumptions.

In the first essay we examine how the content and structure of online reviews of credence services systematically differ from the content and structure of reviews of experience services and how consumers judge these differences. We have found that online reviews of credence services have either less important or less credible content

than reviews of experience services and that consumers do discount the credibility of credence claims. However, while consumers rationally discount the credibility of simple credence claims in a review, more complex argument structure and the inclusion of evidence attenuate this effect.

In the second essay we ask, “Can online reviews predict the worst doctors?” We examine the power of online reviews to detect low quality, as measured by state medical board sanctions. We find that online reviews are somewhat predictive of a doctor’s suitability to practice medicine; however, not all the data are useful. Numerical or star ratings provide the strongest quality signal; user-submitted text provides some signal but is subsumed almost completely by ratings. Of the ratings variables in our dataset, we find that punctuality, rather than knowledge, is the strongest predictor of medical board sanctions. These results challenge the definition of credence products, which is a long-standing construct in economics of information theory. Our results also have implications for online review users, review platforms, and for the use of predictive modeling in the context of information systems research.

ASSESSING QUALITY IN HIGH-UNCERTAINTY MARKETS: ONLINE REVIEWS OF CREDENCE SERVICES

by

Shannon Rose Lantzy

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Associate Professor Katherine Stewart, Chair
Associate Professor Siva Viswanathan, Co-Chair
Research Associate Professor, Joseph P. Bailey
Professor Rebecca Hamilton
Associate Professor Erkut Ozbay

Dedication

Dedicated to my husband, Jared

Acknowledgements

Many heartfelt thanks to my advisors, Kate and Siva, for sticking with me through this long process, and teaching me every step of the way. Same to my co-authors, Rebecca and David, for their collaboration, time, and valuable insights. I also deeply appreciate the support of my committee, department, colleagues, and classmates throughout my tenure at Smith.

Thanks to my village of family, friends, children, and my children's godparents, who fill life with joy, and who make its richness possible.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

Despite living in the information age, consumers of some types of products[1] still face a significant, persistent, and harmful information deficit. The average consumer cannot easily discern whether a doctor's diagnosis is correct or the prescribed treatment is appropriate, and they cannot assess a doctor's skill in treatment. Doctors' services involve such specialized knowledge and individualized care that it is extremely difficult for even third-party verification to assess the quality of an individual doctor (Scholle et al. 2009). Fraud, in the form of overtreatment and overcharging, is clearly evident in the market for healthcare services in the United States. For example, the over-provision of cesarean birth over vaginal delivery is rampant in the US (Amnesty International 2010; Gruber et al. 1999; Rehavi and Johnson 2013). A number of studies establish that fraud occurs in other markets in which the average consumer is unable to assess the quality or necessity of a given service: for example, foreign visitors are frequently subjected to longer and higher-priced taxi rides than locals[2] (Balafoutas, Beck, Kerschbamer, & Sutter, 2011); as many as half of recommended automotive repairs are actually unnecessary (Rasch & Waibel, 2012; Schneider, 2012); and typical consumers receive more unnecessary surgeries, and pay more for them, than more informed patients (e.g., doctors)

[1] We use the word "products" whereas some of the literature refers to "goods." We feel that "products" is the more general term, inclusive of both goods and services.
[2] We note that Uber and other GPS-enabled services may change the taxi services landscape.

(Dulleck & Kerschbamer, 2006; Gruber, Kim, & Mayzlin, 1999; Rehavi & Johnson, 2013). This dissertation examines online reviews as a potential source of new information that may attenuate these harmful information deficits.

The economics of information literature posits that a product can be categorized according to the level and timing of information deficit between buyers and sellers in its market. Search, experience, and credence products are differentiated by when a consumer can reduce her information deficit and how much effort is required to do so (Darby & Karni, 1973; Nelson, 1970). For search products, quality information is easy to find and verify through inspection prior to purchase. For experience products, quality information cannot be reasonably obtained prior to consumption. For example, one can only reasonably assess the quality of a can of tuna after having a taste (Nelson 1970).

Credence products, such as those cited in the opening paragraph, often leave consumers with persistent information deficits even after consumption. In their original paper defining the concept, Darby & Karni (1973) used automobile repair services as the exemplary credence product. The average consumer is not an expert in automobile mechanics. To have an automobile repaired, the consumer must trust that the mechanic's diagnosis and proposed repairs are necessary and well-done. Because the profit-seeking mechanic knows the consumer cannot verify his diagnosis and repair, the mechanic has a strong incentive to overcharge or overdiagnose his customer with little risk of recourse (Pesendorfer and Wolinsky 2003). Thus, auto repair services are credence products. Other credence products include healthcare

services, taxi rides in foreign countries, organic or fair-trade foods, or age-defying facial creams (Balafoutas et al. 2013; Hsieh et al. 2005; Mccluskey 2000). For each of these products, it is practically impossible for consumers to assess the core quality of the product and to know whether they are being duped into unnecessary or overly expensive services.

Most products exhibit multiple qualities (or multiple "attributes," as we will say in the remainder of this dissertation). For example, although a consumer might be able to evaluate the prices of entrees (a search attribute) prior to eating at a restaurant, it is very difficult to evaluate their tastiness or the restaurant's service (experience attributes) without having consumed a meal. And, even after eating the meal, a consumer cannot verify the claim that it was made with organic ingredients (a credence attribute). Because restaurants are primarily selected for their food and service quality, we suggest restaurants can be classified as experience services. Healthcare services and auto shops, in contrast, are usually chosen for service provider attributes such as knowledge and skill in diagnosis, which are credence attributes. Doctors and mechanics are therefore frequently referenced in the literature as examples of credence service providers.

As mentioned above, economic theory predicts that consumers of credence services are vulnerable to a wide range of fraudulent and unethical activity, and empirical evidence supports that prediction (Dulleck & Kerschbamer, 2006; Wolinsky, 1993). This consumer vulnerability arises from information asymmetry; in other words, the providers of products and services have access to information about

product or service quality that is to some degree inaccessible by the consumer. In a theoretical market with pure information asymmetry, the market fails (Akerlof 1970). There are a number of proposed solutions to prevent market failure, such as government regulation (e.g., licensure), expert third-party evaluations (e.g., Consumer Reports), warranties, and reputation-building. Each of these mechanisms attenuates, but does not resolve, the consumer's information deficit (Ely and Valimaki 2003; Hahn and Hird 1991; Joskow and Rose 1989). New mechanisms that supply credence attribute information in markets for credence products have the potential to reduce the burden of fraud borne by consumers.

Online reviews are relatively new, and their potential to supply credence attribute information has not been critically assessed. Theoretically, since online reviews allow consumers to share information about their service experiences and evaluations with other consumers, they should reduce market information deficits between potential consumers and service providers. Huang, Lurie, & Mitra (2009) demonstrated that as a result of consumers posting their experiences online, products that previously required trial and sampling before purchase (i.e., experience products) have begun to behave more like search products. In this case, online reviews filled an information deficit in the market for experience products. However, because, by definition, consumers cannot assess the quality of credence attributes, consumer reviews of credence services (e.g., doctors, auto mechanics) are of dubious credibility and usefulness. Indeed, many doctors have strongly rejected the legitimacy of consumer reviews, noting that consumers are not technically equipped to evaluate

their services (Andrews, 2008; Jain, 2010). Some doctors have been so concerned about the potential effects of online reviews that they have asked patients to sign documents promising never to review their doctor (ElBoghdady, 2012). Nonetheless, consumers utilize numerous forums to review credence services (e.g., RateMDs.com, Angie's List, and Yelp).

Yet despite this contrast between increasing utilization and service provider skepticism of online reviews for credence services, extant research has focused primarily on reviews of search and experience products (Huang, Lurie, & Mitra, 2009; Mudambi & Schuff, 2010; Park & Kim, 2008). What little information exists about credence service reviews does not address the question of whether online reviews are actually useful in informing patients about credence attributes. Research on doctor reviews has focused on comparing online reviews to experience information. Namely, Gao et al. (2011) compare online reviews to patient experience ratings (experience information). Lu and Rui (2015) examine the relationship between online reviews and immediate mortality rates (which are observable immediately after treatment). Wallace et al. (2014) examine associations between online reviews and state-level data such as mortality and 14-day readmission rates (both observable immediately after treatment).

In addition to measuring whether online reviews could supply new information to the market, we want to understand whether and how online reviews of credence services differ from online reviews of experience services, and we inquire how these differences may affect consumers. There is a wide range of research on the

impact of online reviews for products like movies and restaurants. There is a natural "fit" between movies, books, and meals and their reviews: consumers write online reviews about their *ex post* evaluation after experiencing a movie or a restaurant, so these reviews provide exactly the type of information that *ex ante* consumers need to evaluate for their purchase decision. There is no such fit between credence services and credence service reviews. While credence services have some experience attributes (e.g., a doctor's bedside manner), they are defined by the dominance of their credence attributes. It is unclear whether a consumer's *ex post* evaluations of a credence service can contain any information about credence attributes. Credence reviews could simply contain mentions of consumer experiences, which by definition would be solely experience attribute information. Alternatively, consumers may speculate about credence attributes, or consumers with expert knowledge could offer their expert insight or opinions regarding credence attributes. It is unclear which of these might be true, and therefore it is unclear whether online reviews can be useful in evaluating credence attributes and services.

Though online reviews do not immediately seem like a fitting solution to the credence attribute information deficit, there is one pair of studies that, together, provide evidence that perhaps online reviews can help to fill this deficit in a way that was previously done only by government agencies. Jin and Leslie (2003) demonstrated that government-supplied information has a significant impact on reducing hygiene-related illness caused by restaurants. In a paper one decade later, Kang et al. (2013) demonstrated that online reviews carry a useful signal of restaurant

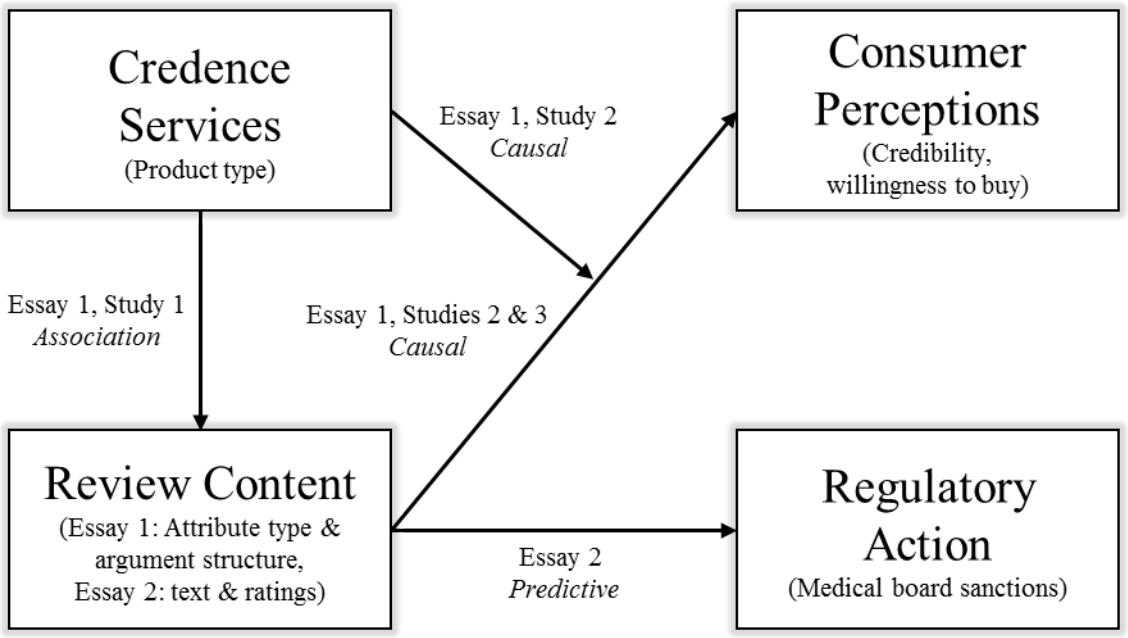
hygiene practices, which are typically a credence attribute. These two papers, taken together, make a strong analogous case for the value of inquiry into online reviews and their potential to supply credence attribute information.

We investigate the broad question, “Are online reviews a useful and credible source of information for consumers of credence products?” In Essay 1, we answer the following: 1) Does the content or structure of reviews of credence services on forums such as Yelp.com differ from the content or structure of reviews of experience services, and if so, how? 2) Which attributes (search, experience or credence) do consumers believe are most important for evaluating credence service providers? 3) How do the content and structure of reviews influence the extent to which consumers find the reviews credible and their willingness to purchase the reviewed service? In our first study, we analyzed the content and structure of real online reviews of both credence and experience services, then conducted a series of two related experimental studies designed to answer these questions. The first two studies focus on the type of attributes (experience or credence) discussed in reviews, and on consumer response to those attributes: the first study provides a content analysis of real reviews, and the second offers an experiment to measure how attribute type affects consumer perceptions regarding the information’s credibility. The third study was designed to test the credibility of combinations of credence claims and evidence, experience evidence, and consumers’ willingness to purchase a credence service based on the reviews.

The second essay examines usefulness of online reviews for a specific credence product, doctor services, when compared against government regulatory information. We ask the following questions: 1) How do online reviews compare with institutional quality disclosure mechanisms? 2) More specifically, what characteristics of online reviews serve to provide a predictive signal of suitability to practice medicine? 3) Can online reviews predict institutional judgments of low quality? To answer these questions we use online doctor reviews to build predictive models of institutional quality disclosure – specifically, state medical board sanctions. Data for the second essay was collected from RateMDs.com (with permission) and the Federation of State Medical Boards (in collaboration). We linked these two databases, prepared variables based on online ratings and their textual reviews, built predictive models of the reviews, and evaluated model performance.

A conceptual model of our two essays is presented in Figure 5.

Figure 1: Conceptual model of research



Results from the first essay demonstrate that online reviews of credence services significantly differ from experience service reviews in their content and structure. Credence service reviews include more mentions of credence attributes, and credence claims appear in a higher proportion than evidence used to support the claims. This suggests that online reviews for credence services do not hold the same kind of information as reviews of experience services. The first experimental study suggests that consumers' perceptions of credence service reviews and credence attribute claims are significantly different from their perceptions of experience service reviews and claims; consumers are more skeptical of credence claims. We find that the perceived credibility of reviews is sensitive to differences in content (i.e., discussion of credence vs. experience attributes) and structure (i.e., supported vs.

unsupported claims). Consumers rationally discount the credibility of simple credence claims in a review, but more complex argument structure or the inclusion of evidence attenuate this effect. There are several important implications of these differences in content and consumer perceptions. First, broad conclusions derived from the existing online review literature may not extend to credence service reviews. For example, the established finding that higher ratings lead to higher sales may not apply (Chevalier & Mayzlin, 2006). Furthermore, information platforms such as Yelp and Angie's List do not differentiate between credence and experience services or attributes in the design of their review systems. These platforms may be able to improve their product by designing service review forms and templates tailored by product type.

The second essay demonstrates that doctors who will be sanctioned have lower average ratings, higher variance, and a higher volume of reviews before their sanctions than their never-sanctioned counterparts. These features, combined with demographic information, provide predictive power of a doctor's unsuitability to practice medicine. In other words, we show that online reviews are useful for consumers who want to avoid low-quality doctors. Surprisingly, textual features within the review do not substantially add predictive power, suggesting that while consumers may be able to perceive their doctor's quality, they do not articulate those differences using distinct words in text.

CHAPTER 2: ESSAY 1 – ONLINE REVIEWS OF CREDENCE SERVICES: AN ANALYSIS OF THEIR CONTENT, STRUCTURE, AND PERCEIVED CREDIBILITY

Motivation

Online reviews allow consumers to share information about their service experiences with other consumers. This post-purchase information-sharing may reduce the information deficit of prospective consumers because many attributes of a service experience cannot be evaluated prior to consumption (i.e., they are credence or experience attributes rather than search attributes (Huang et al. 2009)). For example, although a consumer might be able to evaluate the location of a doctor's office (a search attribute) prior to visiting, it is very difficult to evaluate the doctor's bedside manner (an experience attribute) without having had an appointment. Even after an appointment, the average consumer cannot evaluate the knowledge of her doctor (a credence attribute). This raises an important issue: given that consumers cannot assess the quality of credence attributes, can they provide relevant information to other consumers after using services such as doctors and auto mechanics, which are dominated by credence attributes? The literature suggests that some doctors believe that patients cannot make a competent written assessment of their doctor's knowledge level (Andrews 2008; Jain 2010), and in some cases patients are even required to sign documents promising never to review their doctors (ElBoghdady 2012).

Surveys of consumers show that online reviews of doctors are increasing in

number and utilization (Fox and Jones 2009). Although consumers have numerous forums to review credence services (e.g., RateMDs.com, Angie's List, Yelp.com), extant research considering product type has focused on reviews of search and experience products (Huang et al. 2009; Mudambi and Schuff 2010; Park and Kim 2008). As noted above, while there is a clear fit between experience products and experience reviews, it is unclear how credence products and reviews "fit." Furthermore, we do not know how consumers reading the reviews will process them. We address these gaps by examining the content and structure of credence service provider reviews and investigating how these characteristics influence consumers' perceptions of the reviews.

We pose three interrelated research questions: 1) Does the content or structure of reviews of credence services on forums such as Yelp.com differ from the content or structure of reviews of experience services, and if so, how? 2) Which attributes (experience or credence) do consumers believe are most important for evaluating credence service providers? 3) How do the content and structure of reviews influence the extent to which consumers find the reviews credible and their willingness to purchase the reviewed service?

We address these questions by conducting a content analysis of real online reviews and a series of lab experiments. First, we content analyze online reviews of service providers to compare their content and structure across services that are dominated by credence attributes. Each review was divided into discrete phrases, which were coded for structure (evidence, claims or other components of an argument

based on Toulmin's 1958 framework) as well as specific service provider attributes mentioned in reviews. Results indicate that reviews of credence services include claims about credence, experience and search attributes, whereas reviews of experience services rarely include credence claims. Further, reviews were systematically less likely to contain evidence related to credence attributes than to experience attributes, suggesting that the credence mentions were not necessarily written by experts with special knowledge.

Next, we conducted a series of experimental studies to investigate consumers' perceptions of reviews, manipulating the type of service provider (experience or credence), the type of attribute(s) mentioned in the review (experience or credence), and the quality of argument in a review. These studies show that consumers are capable of perceiving differences in credibility across review types and across service provider types. For example, they tend to be skeptical of credence claims, perceiving them as less credible than experience claims. However, argument quality also matters: consumers find claims supported by evidence to be more credible than claims presented alone.

Credence Services, Credence Attributes

According to economics of information theory, products are distinguished by the time and cost required for consumers to evaluate the product's qualities. Nelson (1970) was the first to differentiate between product qualities that may be evaluated by the consumer before purchase (i.e., search qualities or attributes), and qualities that

can only be evaluated after purchase (i.e., experience qualities). In this literature, products whose overall quality is dominated by search qualities (or "attributes") are classified as search products, and products dominated by experience attributes are classified as experience products. Darby and Karni (1973) extended this framework to "credence qualities which are expensive to judge even after purchase" (p. 69, Darby and Karni 1973). Credence qualities include those that are hard to verify (such as whether a fruit in the store was organically grown) as well as those that are hard to measure (such as a doctor's skill in diagnosis).

Darby and Karni initiated a rich stream of theoretical and empirical literature on credence product markets. The economics literature has largely focused on theoretical ramifications of the steep information asymmetry in credence markets, namely fraud (i.e., overtreatment, overcharging, and under-treatment; e.g., Balafoutas et al. 2011; Beck et al. 2010; Dulleck and Kerschbamer 2006; Emons 1997; Kerschbamer et al. 2009; Liu 2011; Mimra et al. 2012; Wolinsky 1993). The marketing and information systems literatures have focused on how sellers can overcome consumers' lack of information about product quality, e.g., through branding or other marketing strategies (e.g., Bloom and Pailin 1995; Galetzka et al. 2006; Lim and Chung 2011; Srinivasan and Till 2002). We extend this work by examining word of mouth communication among consumers about credence services. What do consumers say about credence service providers when they write reviews, and how do other consumers evaluate this information?

Reviews of Credence Services

As we note above, products and services are a bundle of search, experience, and credence attributes (Darby and Karni 1973; Ford et al. 1990; Lim and Chung 2011; Srinivasan and Till 2002), and products and services may be classified as search, experience, or credence (SEC) based on their most important attributes, i.e., credence attributes are the most important qualities to evaluate in credence services, and experience attributes are the most important qualities to evaluate in experience services (cf. Darby and Karni 1973; Huang et al. 2009; Lim and Chung 2011).

There is a growing body of research that examines distinct product attributes within online word of mouth. For example, Hamilton et al. (2015) investigate how mentions of specific attributes in discussions influence subsequent mentions of those attributes. Decker and Trusov (2010) provide a review of methods for extracting attribute sets and measuring consumer preferences from online reviews and word of mouth. None of the reviewed papers, however, considers the attributes in light of the *ex ante* likelihood of consumers to have access to that information, i.e., whether the attributes are search, experience, or credence.

While online review research does not consider SEC attribute type, there is some work that considers search product versus experience product types (Hao et al. 2010; Huang et al. 2009; Jiménez and Mendoza 2013; Mudambi and Schuff 2010). Huang et al. (2009) demonstrated that online reviews and internet searching have "moved" products that were traditionally dominated by experience attributes into the search product classification. The study explains this phenomenon by suggesting that

attribute evaluations that consumers previously had to make for themselves (experience attributes) can now be approximated by reading other consumers' reviews. We aim to build on the existing literature by examining the composition of online reviews through the lens of multiple product and attribute types: we examine search, experience, and credence attributes mentioned within online reviews of credence and experience service providers.

Very little research has examined reviews of credence products and services. While there is a small body of research on the narrow domain of doctor reviews, this work examines doctor reviews in detail and does not compare doctors as credence services against other product types (Gao et al. 2011; Lu and Rui 2015; Wallace et al. 2014). Given the impact of healthcare on both individual consumers and government regulation, these in-depth examinations can have a huge impact. However, they cannot draw conclusions across product types. Our research bridges a chasm between the large body of inquiry into search or experience product reviews and the smaller, independent body of inquiry into credence product reviews.

We want to understand whether and how online reviews of credence services differ from online reviews of experience services. If experience attributes are most important for evaluating experience services, we expect online reviews to contain information about experience attributes. We cannot expect an analogous relationship for credence services, attributes, and reviews. Consumers are motivated to write reviews that will provide helpful information to other consumers (e.g., Bateman et al. 2006; Hamilton et al. 2015; Hennig-Thurau et al. 2004; Kraut and Resnick 2010; Moe

and Trusov 2011), therefore we suggest that the most helpful information is information about the most important attributes of the reviewed product. This logic, taken alone, leads to the conclusion that online reviews of credence products will mention credence attributes. However, credence attributes are hard for consumers to evaluate, which means they may have little or no information to share. A helpful review writer who does not possess expert knowledge may not want to speculate on credence attribute information. Because we do not know whether consumers will try to share credence attribute information, we cannot predict whether such information will dominate online reviews of credence products. We can, however, suggest with confidence that online reviews of experience products will be filled with experience information, since it is both important and consumers can evaluate it. Thus we can make a prediction about the comparative composition of credence attribute mentions in credence versus experience reviews. We expect credence service provider reviews to contain more mentions of credence attributes than reviews of experience service providers.

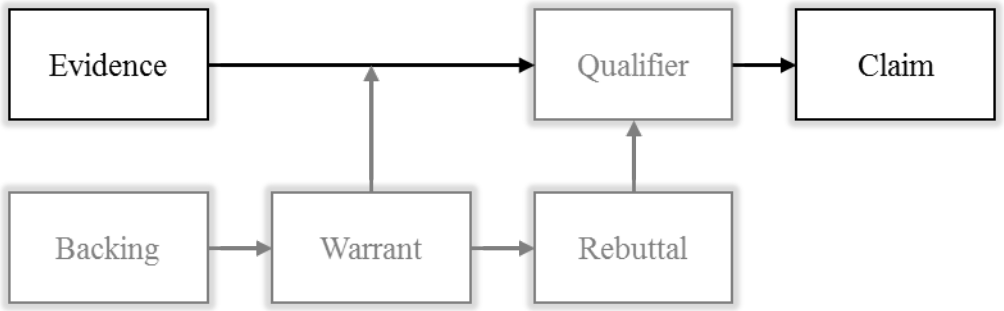
H1: Reviews of credence service providers will mention more credence attributes than reviews of experience service providers.

Extant research suggests that the credibility of an online review depends on both the structure and content of information in the review. Cheung et al. (2009) found that the sidedness (i.e., the balance between positive and negative information) of online word of mouth communications significantly influences consumers' perceptions of review credibility. Nelson (1970) proposed and Ford et al. (1990)

demonstrated that messages with more objective information than subjective information are more credible. Bringing these two streams of research together, we propose that online reviews can be analyzed as arguments and decomposed to assess their argument structure (Kim and Benbasat 2006; Racherla et al. 2012).

We use the classic Toulmin (1958) framework to analyze the components of arguments in reviews. Toulmin's framework includes six types of components: claims (the message's conclusion), grounds or evidence (data to support the claim), warrants (logical link between grounds and claim), backing (supports the warrant), rebuttals (reasonable restrictions on the claim), and qualifiers (words that modulate the degree of certainty of the claim) (see Figure 2). In our studies we focus on claims and evidence, which represent 99% of the information in online reviews, according to our analysis in Study 1. A claim is an assertion, such as "this doctor was prompt," whereas evidence is information that would support the claim, such as "this doctor arrived for my appointment three minutes before the scheduled time."

Figure 2: Toulmin model of argument



Note: Our research focuses on claims and evidence, which represent 99% of the information in our dataset

Although consumers are likely to have evidence about experience attributes after consuming a service (e.g., how many minutes their doctor is early or late for the appointment), they are by definition less likely to have evidence about credence attributes after consuming a service (e.g., whether the doctor possessed critical and up-to-date knowledge of a particular disease). Therefore, when consumers mention experience attributes in a review, they may offer evidence about that attribute, but when they mention credence attributes, they are not likely to offer evidence. We derive the following hypothesis:

H2: Evidence about credence attributes is less likely to be mentioned than evidence about experience attributes.

We conducted a content analysis of online reviews to test our hypotheses.

Study 1: Content Analysis of Online Reviews

In this study, we content-analyzed reviews of five different types of service providers, two of which sell credence services and three of which sell experience services. In addition to explicitly testing our hypotheses, our goal was to understand whether and how reviews of credence services differ from reviews of experience services, investigate the nature of credence attribute information included in credence service reviews, and inquire into the existence of evidence for credence and experience claims.

Method

We collected 158 online reviews from Yelp.com by randomly sampling about 30 reviews for each of five different service providers. Two were credence service providers that have been examined in previous work on credence markets (doctors and mechanics), and three were experience service providers (hair stylists, masseuses, and house cleaners).

Individual reviews typically contain mentions of a number of different attributes and may be structured to include claims, evidence, or both. In order to reliably code the reviews, we divided the 158 reviews into 1,706 mutually exclusive snippets of text (see Table 1 for an example). Next, each snippet was coded into one of 23 attribute codes (Table 2) and one of the structure codes, or into the "other" category (about 32% of the snippets; see Appendix). We developed both the attribute codes and the "other" categories iteratively using an initial training set, then re-coded the full set of reviews after the categories had been established. A subset (40%) of the coding was done by two independent coders to ensure the reliability of the structure and attribute coding schema. Reliability was computed using Rust and Cooil's PRL scores (Rust and Cooil 1994) and was 0.79 for structure codes and 0.83 for search, experience, and credence attributes, which is satisfactory for this type of nascent work (Rust and Cooil 1994).

Table 1: Sample review, division into snippets and coding

Sample review	**Snippets**	**Structure code, Attribute code**
Everyone needs a "car guy" and for me, David at DP is that guy. He is very knowledgeable and will give you nothing but honesty. He uses modern technology (imagine a computer in an auto shop?!) to look for other instances of your problem and potential recalls/safety histories.	1. Everyone needs a "car guy" and for me, David at DP is that guy. 2. He is very knowledgeable 3. and will give you nothing but honesty. 4. He uses modern technology (imagine a computer in an auto shop?!) 5. to look for other instances of your problem and potential recalls/safety histories.	1. Other 2. Claim, Knowledge 3. Claim, Trustworthiness 4. Evidence, Physical space 5. Evidence, Thoroughness

Table 2: Attribute codes

Attribute	**Type**	**Examples**	**Total Mentions**	**Mentions in Cred. Reviews**	**Mentions in Exper. Reviews**
Location	Search	Claim: "good location." Evidence: "it's an easy walk to the Silver Spring Metro."	27	16	11
Reputation	Search	Claim: "she is highly recommended." Evidence: "we called them in based on yelp reviews"	46	14	32
Qualifications	Search	Claim: "I can't imagine any doctor in the area who matches up to her qualifications." Evidence: "she's not Deva trained, but she has worked in a Deva-based salon."	5	1	4
Accuracy of estimate	Exper	Claim: "They do what they say they are going to do" Evidence: "The final cost when I went to pick up my car was $650 more than estimated."	9	6	3
Carefulness	Exper	Claim: "[name] did our home's exterior with care." Evidence: "No one asked me about my health history or that of my family."	81	29	52
Cleanliness	Exper	Claim: "the studio was very clean." Evidence: "the nurse who did my blood work and vaccines did not wear gloves."	13	3	10
Communication skills	Exper	Claim: "He is a great listener." Evidence: "He has gone out of his way to call me with test results."	84	41	43
Customer service	Exper	Claim: "Great service!" Evidence: "they happily changed my oil while my car was in for body work."	58	28	30
Ease of scheduling	Exper	Claim: "Good luck getting an appointment." Evidence: "The office was able to see me on short notice."	71	34	37
Personability	Exper	Claim: "Dr. [name] is very personable" Evidence: "The doctor welcomed me back."	110	63	47
Physical space	Exper	Claim: "The office was beautiful." Evidence: "…you sit in a chair in a comfy room."	45	16	29
Professionalism	Exper	Claim: "Dr. [name] is professional." Evidence: "[name] has since sent me a few nastigrams"	14	3	11
Promptness	Exper	Claim: "the place is a time management nightmare." Evidence: "It took me 2½ hours to get my teeth cleaned."	50	17	33

Attribute	Type	Examples	Total Number of Mentions	Mentions in Credence Reviews	Mentions in Experience Reviews
Staff quality	Exper	Claim: “Their staff is super friendly” Evidence: “he has an in-office nutritionist who has some great ideas”	38	27	11
Value for price	Exper	Claim: “for the price it’s worth it!!” Evidence: “they charge way more than insurance companies allow for out of network care.”	122	27	95
Competence	Cred	Claim: “His diagnoses have always been right on.” Evidence: “Dr. [name] was quick to catch our son’s developmental delays…”	45	16	29
Efficiency	Cred	Claim: “the doctor was efficient.” Evidence: “Car was ready to go the same day.”	52	35	17
Ethics	Cred	Claim: “They operate under numerous names to confuse prospective buyers.” Evidence: “They threatened to break my windows if I left a bad review.”	4	0	4
Honesty	Cred	Claim: “He will give you nothing but honesty.” Evidence: “In addition to never disclosing the water leaks when we were purchasing, they are continuing to show and sell the remaining units without disclosing the leaks.”	14	5	9
Knowledge	Cred	Claim: “His knowledge in the field is exceptional.” Evidence: “After calling a few doctor friends, I learned that the Nurse Practitioner was in fact right.”	23	15	8
Overtreatment	Cred	Claim: “he tends to offer more prescriptions than I need.” Evidence: “I noticed that they had automatically tested me for Hep C without having asked me about previous tests.”	12	10	2
Thoroughness	Cred	Claim: “I have never had a physical that was as cursory as it was with him.” Evidence: “He uses the meticulous notes he made”	4	2	2
Trustworthiness	Cred	Claim: “I love a doctor you can put a lot of trust in.” Evidence: “I don't have a lot of confidence in them not charging my insurance for the $80 test anyway”	15	9	6

The 24 attribute codes were further categorized into search, experience, and credence attributes. Three judges used the definitions of search, experience, and credence from prior work (Darby and Karni 1973; Ford et al. 1988) to independently categorize each attribute. There were no instances in which all three judges disagreed, and disagreements were resolved by discussion. For example, honesty was classified as a credence attribute (Dulleck and Kerschbamer 2006), friendliness as an experience attribute, and location as a search attribute.

Of the six argument structure components in Toulmin's (1958) framework, our coding revealed almost exclusively claims and evidence. Because there were only very rare qualifiers or warrants (5 out of over 1700 snippets) and no examples of the other structure codes, we only discuss claims and evidence in our analysis.

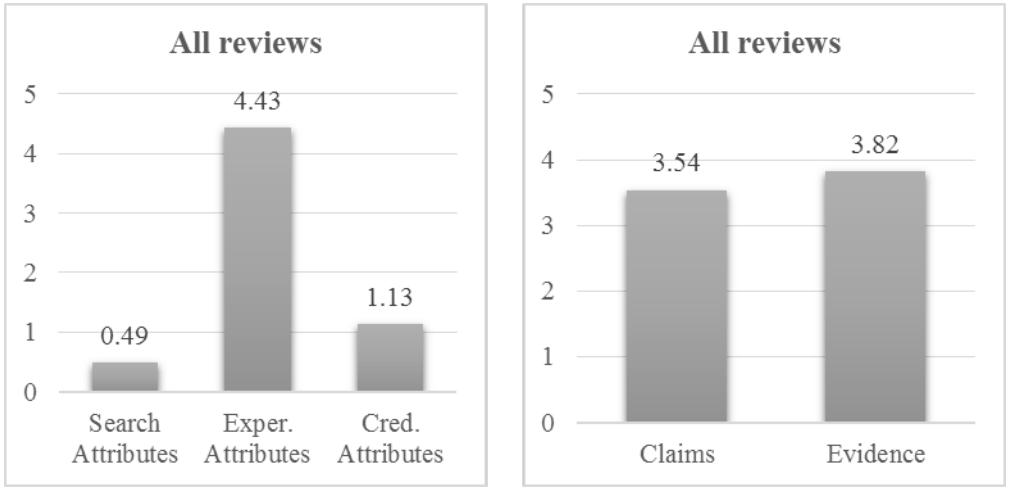
Results

The length of the reviews averaged 845 characters and did not differ for credence (M = 780 characters) and experience service providers (M = 887 characters; $F(1, 156) = .82$, $p > .36$). Most reviews were positive, with rating averaging 4.01 out of 5, and ratings did not significantly differ for credence (M = 4.22) and experience service providers (M = 3.87; $F(1, 156) = 2.52$, $p > .11$). Reviews also did not differ in the number of "useful" votes they received across credence (M = 1.38) and experience service providers (M = 1.55; $F(1, 156) = .19$, $p > .66$).

We counted the number of mentions of each type of attribute code (search, experience, credence, claim, and evidence) in each review. Both experience and

credence service provider reviews mentioned a mix of search, experience, and credence attributes. Overall, there were more mentions of experience attributes per review ($M_{\text{exp attribs}}$ = 4.43) than of credence ($M_{\text{cred attribs}}$ = 1.13) or search ($M_{\text{search attribs}}$ = .49; $F(2, 312) = 167.28$, $p < .001$) attributes. There was approximately the same number of snippets per review classified as evidence and as claims (M_{evidence}= 3.82 versus M_{claims}=3.54; $F(1, 156) = .75$, $p > .38$). Figure 3 shows the share of attribute type mentions and each argument structure type across reviews of all service providers in our sample.

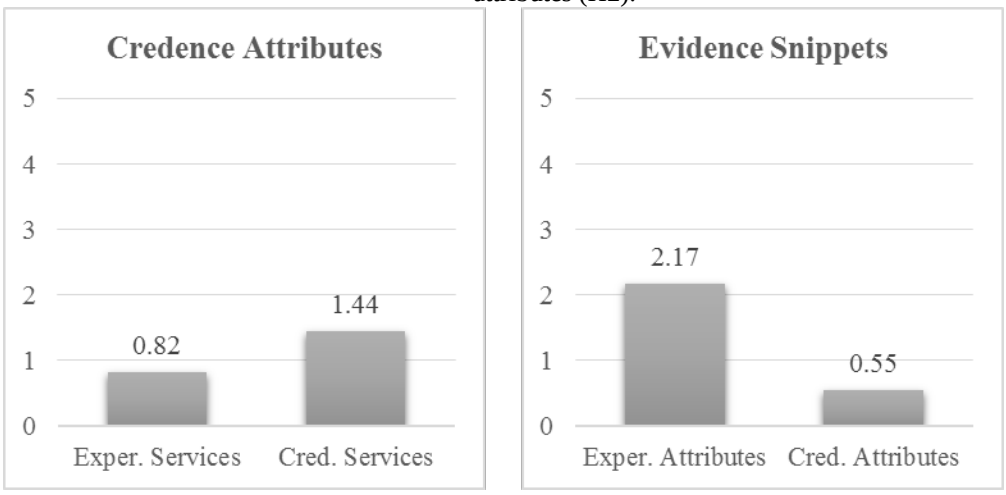
Figure 3: Mean number of attribute type mentions and argument structure types across all reviews



Attributes. We investigated whether reviews of credence service providers mention more credence attributes than reviews of experience providers (H1). As predicted, the count of credence attributes mentioned in reviews of credence providers (M = 1.44) was higher than in reviews of experience providers (M = .82; $F(1, 156) = 7.02$, $p < .01$). Of the 151 reviews mentioning search, experience, and/or credence attributes (96%), the proportion of credence attribute mentions was also

significantly higher for credence (M = 21%) than for experience providers (M = 14%; $F(1, 149) = 4.75$, $p = .03$). Thus, H1 is supported, and shown in Figure 4, left chart.

Figure 4: Left chart: Mean number of credence attribute mentions for experience and credence services (H1). Right chart: Mean number of evidence snippets about experience attributes and credence attributes (H2).



Although there was no difference in the count of experience attributes mentioned in reviews of credence service providers (M = 4.59) and experience service providers (M = 4.27; $F(1, 156) = .36$, $p > .55$), the proportion of experience attribute mentions was marginally higher for experience (M = 79%) than for credence service providers (M = 72%; $F(1, 149) = 3.57$, $p = .06$). There was no difference in the count of search attributes mentioned in reviews of credence providers (M = .48) and experience providers (M = .50; $F(1, 156) = .02$, $p > .90$), and the same was true for proportions (M = 6% vs. 7%, $p > .76$). Table 3 shows the average count of search, experience, and credence attribute type mentions in experience and credence service provider reviews.

Table 3: Comparison of incidence across experience and credence service provider reviews in Study 1

	Experience providers (N = 94)	Credence providers (N = 64)	$F(1, 156)$
Credence attribute mentions^	.82	1.44	7.02**
Credence attribute claims	.36	.81	12.65**
Credence attribute evidence	.44	.59	.86
Experience attribute mentions	4.27	4.59	.36
Experience attribute claims	1.60	1.56	.02
Experience attribute evidence	2.61	2.91	.44
Search attribute mentions	.50	.48	.016
Search attribute claims	.14	.16	.09
Search attribute evidence	.32	.23	.89
Claims	3.32	3.72	1.25
Evidence	3.72	3.90	.08
Summary evaluations	1.37	1.03	3.39
Other	3.89	3.26	2.27
Length of review (characters)	887	780	.82
Rating	3.87	4.22	2.52
Useful votes	1.55	1.38	.19

^ Note: The sum of claims and evidence may be slightly less than total number of mentions; for the sake of clarity, a very small number of warrants or qualifiers were not included in the breakdown.
** Indicates effect significant at $p < .01$; all other effects nonsignificant ($p > .07$).

Structure. Next, we investigated the relationship between structure and attributes to test whether evidence about credence attributes is less likely to be mentioned than evidence about experience attributes (H2). We compared incidence of claims and evidence for credence and experience attributes, controlling for provider type using a 2 structure (claim vs. evidence) x 2 attribute type (credence vs. experience) x 2 provider type GLMM in which structure and attribute type were

repeated factors and provider type was a between-observations factor. Structure, attribute type and their interaction were significant. No other effects, including provider type, were significant. More snippets that mentioned either credence or experience attributes were classified as evidence (M = 1.64) than as claims (M = 1.08; $F(1, 156) = 16.01$, $p < .001$). Evidence snippets were more likely to be classified as experience attributes (M = 2.17) than as credence attributes (M = .55; $F(1, 156) = 133.95$, $p < .001$). Thus, H2 is supported (shown in Figure 4, right chart).

Discussion

In Study 1, we found support for hypotheses 1 and 2. We hypothesized that credence provider reviews would include proportionally more credence attribute mentions than experience provider reviews because credence attributes are the most important and therefore may be perceived as the most helpful to include in reviews. Overall, though, our analysis suggests that online reviews are dominated by experience attributes. Experience attributes are approximately four times more numerous than either search or credence attributes, regardless of provider type. It stands to reason that search attributes would not be mentioned as frequently as experience attributes in online reviews because reliable information about search attributes can, by definition, be found elsewhere. It also makes sense that there are fewer mentions of credence attributes than experience attributes, even in reviews of credence providers, because consumers writing reviews are unlikely to have information about credence attributes even after they consume a service. We draw the

general conclusion that, overall, consumer reviews are dominated by consumer experiences.

We also sought to understand the structure of real online reviews. We found that reviews do not typically contain complex argument structures, but rather consist of mostly evidence and claims about the characteristics of service providers. Overall, there was roughly the same number of snippets classified as evidence and as claims, but significantly more evidence was provided for experience attributes than for credence attributes when controlling for the base rate.

Since we have found that credence claims appear in online reviews, it is natural to ask what consumers do with this information. In the next section, we review the literature and develop hypotheses about consumers' perceptions of credence claims in online reviews.

Consumer Perceptions of Credence Claims

In the economics of information literature, predictions of consumer and seller behavior in markets for credence products rely on the assumption that consumers are aware of their information deficit. For example, diagnosis of auto repair services is a credence attribute (Dulleck and Kerschbamer 2006); economic theory assumes that consumers are aware that an auto repair diagnosis could be false (Rasch and Waibel 2012; Schneider 2009). Further, theory assumes that consumers correctly place a high importance on evaluating whether the diagnosis is correct, which is one of the most important facets of evaluating the service (Darby and Karni 1973; Dulleck and

Kerschbamer 2006). These assumptions are largely untested in the empirical literature. One lab experiment examining the behavior of sellers and consumers in a credence market shows mixed results (Dulleck et al. 2011) and focuses on transaction outcomes rather than on consumer information processing. Thus, we do not know how consumers perceive the relative importance and credibility of credence and experience attributes in credence provider reviews.

Based on the definition of credence service providers as providers for whom credence attributes are dominant, and on our observation in Study 1 that reviews of credence service providers frequently mention credence attributes, we propose that credence attributes will be perceived as more important than other types of attributes when evaluating credence service providers.

H3: Consumers perceive credence attributes as more important for evaluating credence providers than experience attributes.

Even if credence attributes are important for evaluating credence service providers, it is not clear that claims about credence attributes will be perceived as credible, especially as credence attributes were originally defined as those which "cannot be evaluated in normal use" (Darby and Karni 1973, p. 68). Following Darby and Karni's logic, consumers making credence claims in online reviews do not have the knowledge or expertise required to make an accurate evaluation. Predicting that consumers would be more skeptical of experience claims than search claims and more skeptical of credence claims than experience claims, Ford, Smith and Swasy (1990) conducted experiments measuring consumers' skepticism regarding different types of

claims in an advertising context. Their results partially confirmed the economics of information propositions, suggesting that consumers were more skeptical of experience claims than of search claims, but their results failed to reach significance for the comparison of experience and credence claims. They suggest this may have been due to faulty manipulations. We expect consumers to perceive credence claims as less credible than experience claims in online reviews.

H4: Claims about credence attributes will be perceived to be less credible than claims about experience attributes.

One important moderator of the credibility of credence claims may be argument structure. A higher-quality argument within an online review has been shown to increase a consumer's perception of review credibility. For example, Jensen et al. (2013) varied the "sidedness" of an argument and found that reviews that presented two sides of an argument were more credible than reviews that presented only one side. We propose to test a different form of argument quality in line with the Toulmin framework: claims that are supported by relevant evidence will be more credible than claims without evidence.

In Study 1, we observed that reviews contain a mix of claims and evidence. Some credence claims were supported with evidence. There is some empirical support for the notion that evidence is more credible than claims, showing that consumers find more detailed product information credible (Jiménez and Mendoza 2013). Claims require the reader to evaluate the credibility of the claim's source. In other words, readers must assess whether the source has the requisite knowledge,

experience, and evidence to support a conclusion. In contrast, evidence provides factual information and allows the reader to form his own conclusion, thereby avoiding the need to judge whether the source is making appropriate inferences. When a claim is supported by evidence, the basis of the reviewer's claim becomes clear and doubts about credibility may be alleviated. For example, a review stating, "This mechanic was extremely knowledgeable" constitutes a claim with no supporting evidence. "This mechanic was extremely knowledgeable; he knew what was wrong with my car and was able to fix it easily" constitutes a review that contains a claim with evidence to support it. We predict that consumers will perceive this difference in argument structure and perceive reviews as more credible when they contain credence claims supported by evidence than when they contain credence claims alone.

H5: Reviews with credence claims that are supported with evidence will be perceived as more credible than reviews with unsupported claims.

Ultimately, however, it is not the perceived credibility of online reviews that likely concerns service providers, but rather how these reviews may influence consumers' willingness to choose the focal provider. To determine how consumers are likely to respond in the marketplace based on the reviews they read, we measure purchase intentions.

Several studies of online consumer behavior have shown that purchase intentions are influenced by consumers' beliefs about important attributes such as the perceived trustworthiness of the provider (Daniel et al. 2006; Pavlou and Fygenson

2006; Schlosser et al. 2006; Stewart 2003). In general, credible information has a stronger impact on behavior than less credible information (Pornpitakpan 2004). Thus, more credible reviews about important attributes such as trustworthiness should have a stronger impact on purchase intentions than less credible reviews.

We predict, then, that the valence of information will have opposing effects on willingness to choose. While more credible positive arguments in a review should increase consumers' willingness to choose the focal provider, more credible negative arguments should decrease consumers' willingness to choose the focal provider. Formally:

H6a: Consumers' willingness to choose a focal provider will be higher after reading a supported positive credence claim than an unsupported positive credence claim.

H6b: Consumers' willingness to choose a focal provider will be lower after reading a supported negative credence claim than an unsupported negative credence claim.

We conducted two experiments to test our hypotheses. The first experiment tests H3 and H4, and the next tests H5, H6a, and H6b.

Study 2: Consumer Evaluations of Claims

The goal of this study was to test H3 and H4 by comparing consumers' evaluations of experience claims and credence claims in reviews of experience and

credence service providers, while controlling for the content and structure of the reviews.

Method

Three hundred and fifty-six Amazon Mechanical Turk (mTurk) workers (49.4% females, M_{age} = 35.76) participated in the study in exchange for a small payment. Participants were asked to imagine that they had recently moved to a new city and needed to find a service provider. Each participant was randomly assigned to one of six different service provider types, three of which were experience service providers (hair stylist, house painter, and masseuse) and three of which were credence service providers (doctor, mechanic, and financial advisor). Each participant read and provided evaluations of six different reviews for their assigned type of service provider.

We prepared six short online reviews for each service provider type. In each review, we varied only the attribute mentioned and the fictional service provider name (e.g., James, David, Richard). Three of the reviews mentioned experience attributes (communication skills, personability, and ease of scheduling an appointment) while three reviews mentioned credence attributes (knowledge, trustworthiness, and intelligence). Each participant saw one review mentioning each attribute (e.g., one review for a hair stylist mentioned communication skills, another mentioned personability, a third mentioned knowledge). We also controlled for review valence so that each participant saw two positive reviews and one negative

review for credence attributes and two positive reviews and one negative review for experience attributes. We deliberately chose several examples of each provider and attribute type, called "replicates," to increase the generalizability of the results.

Thus, the study was a 2 (service provider type: credence or experience) x 3 (provider replicates) x 2 (attribute type: credence or experience) x 3 (attribute replicates) x 2 (valence) mixed experimental design in which service provider type and service provider replicates were manipulated between subjects and in which attribute type, attribute replicates, and valence were manipulated within subjects.

Participants rated the credibility of each review (dependable, honest, reliable, sincere, and trustworthy) and each reviewer (an expert, experienced, knowledgeable, and qualified) using a 5-point scale (Ohanian 1990). We also asked participants to indicate how helpful and useful the review would be in selecting a service provider (Kempf and Smith 1998). After evaluating the six reviews, participants rated the importance of each type of attribute for the focal service provider. Finally, they completed a series of control measures including overall propensity to trust (Mayer and Davis 1999), frequency of using the service provider, familiarity with the service provider (Hamilton and Thompson 2007), use of online reviews, and demographics.

Results

Scale reliability ranged from 0.69 to 0.95. Factor analysis confirmed that our scale items loaded onto the correct constructs (Straub et al. 2004). Credibility, helpfulness and usefulness of reviews loaded onto the same component, suggesting

that these measures tap the same underlying construct. The results were very similar for these three variables, and we focus on review credibility in our analysis.

Importance of Attributes. We analyzed differences in the importance of attributes to test whether credence attributes were perceived to be more important than experience attributes for credence providers (H3). A 2 (provider type) x 2 (attribute type) x 2 (valence) linear mixed model on importance ratings, controlling for provider replicate and attribute replicate, showed main effects of service provider type ($F(1, 2053) = 33.45$, $p < .001$), attribute type ($F(1, 2042) = 77.00$, $p < .001$) and valence ($F(1, 2053) = 9.76$, $p < .01$). In general, credence attributes were perceived to be more important ($M = 3.99$) than experience attributes ($M = 3.57$). Attributes were perceived to be more important when they described credence providers ($M = 3.92$) than experience providers ($M = 3.64$) and when claims were negative ($M = 3.85$) rather than positive ($M = 3.71$). We also observed the predicted interaction between provider type and attribute type ($F(1, 1934) = 34.87$, $p < .001$): credence attributes are perceived to be more important for credence providers ($M = 4.25$) than for experience providers ($M = 3.72$), while experience attributes are perceived to be equally important for credence providers ($M = 3.58$) and experience providers ($M = 3.56$). Attribute replicate was a significant covariate, but when we ran the same analysis for each replicate we observed a very similar pattern of effects. No other effects were significant ($p > .23$).

Review Credibility. A 2 (provider type) x 2 (attribute type) x 2 (valence) linear mixed model on review credibility, controlling for provider replicate, attribute

replicate and attribute importance, showed significant main effects of service provider type ($F(1, 2138) = 15.26$, $p < .001$), attribute type ($F(1, 2109) = 12.29$, $p < .001$) and valence ($F(1, 2129) = 17.72$, $p < .001$). Reviews of experience service providers ($M = 3.52$) were perceived to be more credible than reviews of credence service providers ($M = 3.37$), and positive reviews ($M = 3.52$) were perceived to be more credible than negative reviews ($M = 3.36$). Supporting H4, claims about experience attributes were perceived to be more credible ($M = 3.51$) than claims about credence attributes ($M = 3.38$). Attribute replicate and attribute importance were significant covariates, but when we ran the same analysis for each replicate we observed a very similar pattern of effects. No other effects, including the interaction between service provider type and attribute type ($p > .52$), were significant. Table 4 shows the results of the models of review credibility with and without interaction effects.

Table 4: Perceived credibility of claims in Study 2

Source	Model 1	Model 2
Intercept	1210.19**	812.00**
Credence Attribute^	5.00*	12.29**
Credence Provider	42.43**	15.26**
Negative Valence	39.43**	17.72**
Credence Attribute * Credence Provider		.40
Credence Attribute * Negative Valence		.42
Credence Provider * Negative Valence		2.68
Credence Attribute * Credence Provider * Negative Valence		1.19
Provider Replicate	3.01	3.30
Attribute Replicate	19.43**	8.71*
Attribute Importance	268.44**	184.61**

^Note: Effects are negative, e.g., credence attributes are perceived as less credible than experience attributes. ** Indicates effect significant at $p < .01$. * indicates effect significant at $p < .05$.

Discussion

We found that for credence providers, credence attributes are perceived to be more important than experience attributes (H3). We also found that credence claims are perceived as less credible (H4), i.e., consumers are more skeptical of credence claims than experience claims. These results partially confirm some of the assumptions of the economics of information theory – that consumers are both aware of the importance of credence information and savvy to the fact that such information is not readily available. Notably, these results show that consumers penalize both reviews of credence services and reviews containing credence attribute claims in their judgments of review credibility.

Since we purposely designed Study 2 so that each review mentioned only one attribute, we were able to carefully identify effects of attribute type and structure. As shown by the real reviews we analyzed in Study 1, however, reviews typically contain multiple claims and evidence about multiple attributes. In Study 3, we create more realistic reviews. We mix claims and evidence, and we include experience attributes together with credence attributes. We test our remaining hypotheses, which posit that consumers will be less skeptical of credence claims when they are supported by evidence. In addition to measuring the credibility of the reviews, we measured behavioral intentions toward the service provider, i.e., consumers' willingness to choose the focal provider based on review content.

Study 3: Consumer Evaluations of Reviews Containing Claims and Evidence

Study 3 was designed to examine consumers' perceptions of more complex reviews containing mentions of multiple attributes and including evidence as well as claims. This provided the opportunity to test our hypothesis about argument structure (H5) in a controlled environment. We examined whether including evidence increases the credibility of a credence claim. We also measured consumers' willingness to choose a provider based on a positive or negative credence claim that was either supported by evidence or unsupported (H6a and H6b).

Method

Three hundred and eighty-eight mTurk workers participated in the study in exchange for a small payment. Participants were asked to imagine that they had recently moved to a new city and needed to find a doctor and a mechanic. Participants read twelve different reviews, six for mechanics and six for doctors. Examples of the reviews are given in Table 5. The order of the service providers was counterbalanced, so that half of the participants read doctor reviews followed by mechanic reviews, and half read the reviews in the opposite order.

Attributes were selected based on a pretest (N = 155) measuring the importance of the 24 attributes we identified in Study 1 for doctors and mechanics. We chose six attributes (three experience and three credence) of comparable (high) importance for the two providers. For example, knowledge was ranked first for

mechanics and second for doctors, so we chose knowledge as one of our credence attributes. We used three replicates for each attribute type (credence: knowledge, trustworthiness, and thoroughness; experience: communication skill, personability, and promptness), and two replicates for credence service provider (mechanic or doctor). We also manipulated whether the valence of the review was positive or negative.

We designed the reviews so that we could test how the content and argument structure of a review influences its perceived credibility. To help distinguish the effects of including supporting evidence in a review from the effects of simple review length, we included some conditions in which the review included "filler" information that did not mention any attribute (e.g., "I had a handful of symptoms, so I made an appointment for last week" or "I visited here about a week ago to have some concerns looked into"). Credence claims were presented alone, with "filler" statements, with experience evidence (that did not support the claim), with credence evidence that supported the claim, or with both evidence and filler statements. Two additional conditions presented experience evidence or credence evidence alone (instead of a claim) with a filler statement. In total, we tested eight different configurations of review (see Table 5 for the full list). Each credence claim paired with experience evidence (structures 3 and 4) was tested with all three experience attributes, resulting in 12 different reviews for each participant.

Table 5: Description of conditions and examples for Study 3

Condition	Cred. claim	Cred. evidence	Exper. evidence	Valence *	Examples of reviews*
1. Credence Claim only	Yes	No	No	Positive	From my view, I would say that this mechanic is definitely knowledgeable.
2. Credence Claim with filler	Yes	No	No	Negative	I went in the other day to get a couple issues looked at and possibly treated. From my view, I would say that this doctor is not very knowledgeable.
3. Credence Claim + Experience Evidence	Yes	No	Yes	Positive	I arrived at my scheduled appointment time and was seen immediately. The other mechanic I saw was never on time. From my view, we would say that this mechanic is definitely knowledgeable.
4. Credence Claim + Experience Evidence with filler	Yes	No	Yes	Negative	I visited here about a week ago to have some concerns looked into. I arrived at my scheduled appointment time and had to wait 35 minutes. My other doctor was always on time. From my view, I would say that this doctor is not very knowledgeable.
5. Experience Evidence with filler	No	Yes	No	Positive	I came in last week to have some things checked by the mechanic. I arrived at my scheduled appointment time and was seen immediately. The other mechanic I saw was never on time.
6. Credence Claim + Credence Evidence	Yes	Yes	No	Positive	He immediately figured out what was wrong and was able to fix it. A different mechanic we went to had no idea. From my view, we would say that this mechanic is definitely knowledgeable.
7. Credence Claim + Credence Evidence with filler	Yes	Yes	No	Negative	I came in last week to have some things checked by the doctor. He could not diagnose the problem. I went to a different doctor, who immediately figured it out. From my view, I would say that this doctor is not very knowledgeable.
8. Credence Evidence with filler	Yes	Yes	No	Negative	Last week I had an appointment to have several health troubles diagnosed. He recommended major surgery. I got a different opinion and ended up fixing my problem with two simple prescriptions.

* Note: Examples in this table alternate between positive and negative valence; all types of reviews were presented with both positive and negative valence. Similarly, examples alternate between mechanics and doctor service providers; all types of reviews were presented for both providers.

As in Study 2, we measured the credibility of each review (dependable, honest, reliable, sincere, and trustworthy) and how helpful/useful the review would be in selecting a service provider (Kempf and Smith 1998). Participants used the same scales used in Study 2, and in addition, they indicated their likelihood of choosing the provider based on the review.

Results

Review credibility. We ran a linear mixed model on review credibility with review structure, valence, and their interaction as predictors and provider replicate and the length of the review (in characters) as covariates. The effects of review structure ($F(7, 1047) = 65.50$, $p < .001$), valence ($F(1, 2367) = 61.90$, $p < .001$) and their interaction ($F(7, 885) = 3.57$, $p = .001$) were significant. Negative reviews were generally perceived to be less credible than positive reviews, but this effect was attenuated when credence claims were combined with credence evidence, with or without filler statements. Length was not a significant covariate ($p > .54$), but the provider covariate was significant ($F(1, 3904) = 3.94$, $p < .05$), indicating that reviews of doctors were less credible than reviews of mechanics. To investigate whether our results were driven by the results for one provider type, we split the sample into doctors-only and mechanics-only samples and ran the same linear mixed effects model and contrasts on each provider-specific sample. Our results hold for each of these sub-samples, indicating that the results are not driven solely by either

provider type.

To test whether credence claims that are supported with evidence are perceived as more credible than unsupported claims (H5), we ran a planned contrast comparing the perceived credibility of reviews that contained only credence claims (with or without filler) with reviews that contained credence evidence (with or without filler) in addition to the credence claim (i.e., structures 1 and 2 versus structures 6 and 7; $t(3182) = 12.40$, $p < .001$, see Table 6). Our contrast revealed that claims supported with evidence are considered more credible, and thus H5 is supported. As a robustness check, we verified that a credence claim supported by credence evidence was more credible than a credence claim paired with unmatching experience evidence, which was confirmed (structures 6 and 7 versus structures 3 and 4; $t(2203) = 15.20$, $p < .001$). Thus, reviews with better arguments are more credible: reviews including credence evidence that "matches" the credence claim are more credible than reviews including experience evidence that does not "match" the claim. However, the inclusion of any evidence at all increases the credibility of the review, whether this evidence matches the claim or not. We found that credence claims paired with experience evidence were more credible than a credence claim alone (structures 1 and 2 vs. structures 6 and 7; $t(3150) = 7.19$, $p < .001$). Surprisingly, when we controlled for the length of the review (in character count), even credence claims paired with filler statements were more credible than credence claims alone (structure 1 vs. structure 2; $t(1955) = 4.83$, $p < .001$).

Willingness to Choose. We ran a linear mixed model on willingness to

choose the focal provider, with review structure, valence and their interaction as predictors and provider type and the length of the review (in characters) as covariates, plus controls. The main effects of review structure ($F(7, 1216) = 9.15$, $p < .001$), valence ($F(1, 2452) = 1177.43$, $p < .001$) and their interaction ($F(7, 752) = 34.17$, $p = .001$) were significant. Willingness to choose a provider was significantly lower for negative than for positive reviews, and the effect of valence was stronger for review structures with supported claims and claims with evidence or filler than for claims alone.

To conduct more focused tests of our hypotheses, we split the sample into a positive-reviews sample and a negative-reviews sample. Results for willingness to choose a focal provider based on positive reviews are similar to the results for review credibility (Table 6). Supporting H6a, reviews that contained positive, supported claims made consumers more willing to choose a focal provider than positive claims without evidence ($t(1543) = 9.10$, $p < .001$) or positive claims paired with unmatched (experience) evidence, ($t(624) = 10.34$, $p < .001$). Claims paired with experience evidence were more convincing than claims alone ($t(1463) = 5.06$, $p < .001$).

For negative reviews, we observed the reverse pattern of effects (Table 6). Supporting H6b, reviews that contained negative, supported claims made consumers less willing to choose a focal provider than negative claims without evidence ($t(1759) = 3.71$, $p < .001$). Although claims paired with experience evidence were more convincing than claims alone ($t(1668) = 2.15$, $p < .05$), consumers were less willing to choose providers after reading supported negative claims than they were after

reading negative claims paired with unmatched (experience) evidence, ($t(551) = 3.33$, $p < .01$).

Table 6: Means comparison of review structures in Study 3

Structure	Description	Review credibility	Willingness to choose provider	
			Positive review	Negative review
1	Credence claim only	2.739[a]	3.039[a]	3.435[a]
2	Credence claim, filler	3.217[b]	4.083[b]	3.174[b]
3	Credence claim, experience evidence	3.492[c]	4.415[c]	2.999[b]
4	Credence claim, experience evidence, filler	3.682[c]	4.801[d]	2.631[c]
5	Experience evidence, filler	3.579[c]	4.252[c]	3.056[b]
6	Credence claim, credence evidence	4.036[d]	5.419[e]	2.878[b]
7	Credence claim, credence evidence, filler	4.046[d]	5.638[e]	2.049[d]
8	No claim, credence evidence, filler	4.034[d]	5.640[e]	2.476[c]

Note: Means in the same column with different superscripts are significantly different, $p < .05$.

Discussion

In this study, we demonstrated that consumers find reviews pairing credence claims with evidence more credible than reviews without evidence. In other words, consumers are less skeptical of credence provider reviews when they include high-quality arguments (as defined by Toulmin) than when they include only claims. Considering the previously-mentioned controversy about online doctor reviews (ElBoghdady 2012), this finding is cause for both optimism and concern. Although it is comforting to observe that our study participants discerned between a well-argued claim supported with evidence and an unsupported claim, our study also showed that *any* information paired with a credence claim tends to increase its perceived credibility. Thus, it is not only high-quality arguments that are rewarded but also low-quality combinations such as credence claims paired with filler statements. This is

consistent with earlier research by Langer et al. (1978), who show that study participants were more likely to comply with requests (e.g., “May I use the Xerox machine?”) when they were accompanied by “placebic” information (e.g., “May I use the Xerox machine because I have to make copies?”). Given the frequency with which we observe credence claims paired with irrelevant evidence in Study 1 (see Appendix), this pattern of effects is disconcerting.

We also measured consumers’ willingness to choose a credence provider to learn whether better structured and hence more credible reviews have a stronger influence on provider choice than less credible reviews. Our results show that as expected, a positive and credible review makes consumers more willing to use the reviewed service provider, while a negative and credible review makes consumers less willing to use the service provider. Notably, we also observe that consumers are more willing to choose a provider when an unsupported claim is paired with irrelevant information than they are when the claim is presented alone.

General Discussion

In this research, our goal was to test whether the content and structure of online reviews of credence service providers like doctors and mechanics differ from those of experience service providers like landscapers and hair stylists, and whether the content and structure of these reviews influences their perceived credibility or consumer behavior. We found that real online reviews of credence service providers on Yelp.com do contain claims about credence attributes, and that claims about

credence attributes appear more frequently in reviews of credence service providers than in reviews of experience service providers (see Table 7 for a summary of the hypotheses and results). When we examined consumer perceptions of credence claims in a series of controlled experiments, we found that consumers discount the credibility of credence claims more than experience claims when these claims are presented in a very simple format. However, as reviews became more complex, the effects became more nuanced. When credence claims were provided with matching credence evidence, the reviews were perceived to be much more credible than unsupported credence claims. But even pairing a credence claim with unrelated evidence about an experience attribute significantly increased the credibility of the review.

Table 7: Results of hypothesis testing

Hypothesis	**Results**
H1: Reviews of credence service providers will mention more credence attributes than reviews of experience service providers.	Supported by Study 1
H2: Evidence about credence attributes is less likely to be mentioned than evidence about experience attributes.	Supported by Study 1
H3: Consumers perceive credence attributes as more important for evaluating credence providers than experience attributes.	Supported by Study 2
H4: Claims about credence attributes will be perceived to be less credible than claims about experience attributes.	Supported by Study 2
H5: Reviews with credence claims that are supported with evidence will be perceived as more credible than reviews with unsupported claims.	Supported by Study 3
H6a: Consumers' willingness to choose a focal provider will be higher after reading a supported positive credence claim than an unsupported positive credence claim.	Supported by Study 3
H6b: Consumers' willingness to choose a focal provider will be lower after reading a supported negative credence claim than an unsupported negative credence claim.	Supported by Study 3

Healthcare providers seem to be justified in their concern that reviews may

hurt their business with claims about credence attributes. Patients can and do review credence attributes in reviews, despite some doctors' attempts to legally limit patients' rights to review them. If a patient writes a negative review containing a credence claim about attributes such as a provider's knowledge or ability to diagnose problems, our results suggest that readers of the review may perceive it as credible even when the credence claim is not supported by credence evidence. Pairing a credence claim with filler or with experience evidence both make a negative credence claim more credible than presenting the claim alone.

This research clearly illustrates the usefulness of thinking about service providers – whether they provide experience or credence services – as a bundle of search, experience and credence attributes. Our analysis of Yelp.com reviews suggests that reviews of both experience and credence providers are dominated by discussion of experience attributes. This is reasonable: discussion of search attributes is less valuable to readers of the reviews because this information is easily available from other sources, while information about credence attributes is hard to obtain. Thus, online word of mouth may be a better fit for experience service providers than it is for credence service providers. However, it is also important to note that experience attributes of credence service providers are not irrelevant; indeed, they may be important to some consumers. For example, bedside manner is a critical attribute for some consumers when choosing some kinds of doctors. Further, experience attributes themselves may be useful if they are correlated with underlying quality. (We investigate this question in more depth in Essay 2.)

Our work underscores the need for more research examining consumer information processing in the context of credence services. Although we carefully examined the content of real reviews, we have not studied when and why reviewers choose to include credence claims in their reviews. We anticipate follow-up research that elucidates why, for example, a reviewer might make a claim without supporting evidence, or why they may include experience evidence instead of credence evidence. It is also important to examine the conditions under which our results hold. In our experimental studies, participants were asked to read reviews, and they were given a limited number of reviews (6 in Study 2, 12 in Study 3). In the course of a real decision making process, consumers might encounter a much higher number of reviews and devote less time to reading each of them. This could exacerbate some of the worrisome effects we observe, such as the tendency to infer that a credence claim is more credible when it is paired with irrelevant information.

In addition to their implications for consumers and service providers, our findings also suggest several implications for online review platforms such as Yelp.com and Angie's List. We suggest that review platforms can play a role in improving the credibility of reviews. For example, rather than eliciting reviews using an unstructured, open-ended format, review platforms could use structured feedback forms that encourage consumers to make high-quality arguments. Review platforms might also elicit information in two steps so that consumers are encouraged to provide evidence to back up any claims they make in their reviews. In order to prompt consumers to thoughtfully evaluate the credibility of the information they are

reading, firms like Yelp or Angie's List might include cautionary advice to review users. Further, platforms could offer a credibility rating for the reviews themselves. For example, Yelp currently offers voting buttons to label a review as "funny," "cool," or "useful." Yelp could add a "credible" or "evidence-based" button so that savvy consumers could highlight a particularly credible review.

CHAPTER 3: ESSAY 2 – CAN CONSUMERS USE ONLINE REVIEWS TO AVOID UNSUITABLE DOCTORS? EVIDENCE FROM RATEMDS.COM AND THE FEDERATION OF STATE MEDICAL BOARDS

Motivation

Physician services are a ubiquitous example of a theoretical credence product because it is hard for a patient to verify a diagnosis or to assess the need for a specific treatment (Dulleck and Kerschbamer 2006). This information deficit leaves the average patient vulnerable to sub-optimal care. Without information, patients cannot evaluate diagnosis and treatment services, and doctors can therefore over-treat, badly treat, or overcharge their patients.

Regulation in the form of medical licensing is intended to attenuate this problem by requiring a minimum level of expertise to obtain a license and by revoking or suspending licenses of unsuitable providers. Medical boards can issue punishments, such as placing doctors on probation or revoking their licenses permanently, to doctors who violate minimal standards of practice. We use these medical board sanctions as a signal that a doctor receiving the sanction is of lower quality or unsuitable for practice. While most consumers would likely prefer to choose a provider who substantially exceeds this minimal quality threshold, avoiding the lowest quality doctors is a worthwhile goal.

Strict credence theory posits that consumers have no market-based sources of information by which to avoid unsuitable doctors, regulators must step in. In order to

provide a backdrop for our investigation, we review the impact of government intervention on the information that is available to consumers.

Online reviews have engendered a plethora of studies, including investigations of the perceived usefulness of online reviews of doctors. To the best of our knowledge, however, there are no investigations of how online reviews measure the information that regulators worry about the most: the quality and safe practices of practitioners they license (see Appendix for a sample of state medical board mission statements, p. 109).

As we found in prior research, online reviews contain a mix of information, including credence claims. But we do not know whether those claims are related to a given doctor's suitability to practice medicine. Theory predicts that online reviews cannot contain information on the suitability of a physician precisely because their suitability to practice medicine is obscure to the consumer. However, this prediction has not been subjected to much empirical research. Is it possible that online reviews carry some signal of doctor suitability? Are they practically useful for consumers who wish to avoid the worst doctors? We investigate these questions by examining the relationship between the largest body of online doctor reviews, RateMDs.com, and medical board sanctions, which are the standard government intervention for low-quality practitioners.

We find that online ratings can be used to predict which doctors will receive a sanction from a state medical board, suggesting that these ratings can carry a signal of low quality. Online reviews, then, may be a valid source of information for

consumers seeking quality information in credence product markets. This finding is perhaps surprising, as it runs directly counter to credence theory. A more nuanced view, however, shows that not all information in reviews is created equal. In our study, numerical ratings and demographics were found to be more useful than the text within unstructured comments.

Regulation as a Solution to Information Deficits

Regulation is intended to mitigate imperfections and failures that lead markets to perform suboptimally (Joskow and Noll 1981; Parker and Kirkpatrick 2012). Imperfections include missing markets, public good problems, monopoly characteristics, anti-trust issues, and information failures, which are the focus of our inquiry in this paper.

According to some theorists, competitive market mechanisms generate sufficient information for adequate market performance even in markets with imperfect information. One common proposition offered to explain market self-regulation is the theory of unraveling (Beales et al. 1981; Dranove and Jin 2010). The intuition behind this theory is simple: in a market where no quality information is available, consumers are only willing to pay for average quality. The highest-quality seller in a market has an incentive to disclose his quality in order to differentiate himself from the rest and garner a higher-than-average price. The consumer is then willing to pay a higher price for that seller but will only pay the average price for the rest of the sellers. The next-highest seller is then incentivized to disclose his quality,

and so on until all sellers disclose (Grossman 1981; Jovanovic 1982). However, unraveling fails for a number of reasons. For instance, disclosure is sometimes costly, sellers do not always have perfect information about their quality, and consumers do not know the distribution of available quality levels (see Dranove and Jin, 2010, for a thorough review of unraveling failures).

Market information failures may also occur due to the nature of the quality information itself. Several examples follow. First, quality information may have public-good properties, such as the health benefits of a common commodity like milk. If one seller produced and released the information, all other sellers of milk would benefit from the information for free. This creates an individual disincentive to generate and contribute the information (Beales et al. 1981). Second, when quality information is not easily verifiable, sellers may be incentivized to produce false claims, such as the organic raising of livestock. Since farming methods are very difficult for consumers to assess directly, claims by the farmer cannot be verified without unbiased third-party assessment. This restricted possibility of consumer assessment has proven to be a central issue in the credence products problem (Darby and Karni 1973; Dranove and Jin 2010; Dulleck et al. 2011). Third, when quality is measured by proxy via noisy signals, sellers may begin to compete on the imperfect signal rather than on the underlying quality it represents (Rothschild and Stiglitz 1992; Spence 1973). For example, job market candidates may overinvest in education, which is a *signal* rather than a *measure* of quality in workers (Spence 1973). In situations where sellers enjoy market power, they may be incentivized to

reduce quality unnecessarily through planned obsolescence or intentional production failure (Salop 1977). All these examples of information failure have been found to warrant government intervention (Beales et al. 1981; Joskow and Noll 1981).

Information is never perfect, however, policymakers must decide when consumer information is insufficient for satisfactory market efficiency and consumer protection. Governments have a number of tools to solve information failures, including licensing (e.g., medical licensing), guarantees (e.g., FDIC), inspection (e.g., restaurant hygiene inspections (Jin and Leslie 2003)), direct quality disclosure (e.g., Dranove et al. (2003)), and tort law (e.g., legal liability (Schwartz and Wilde 1978)). Efficiency in regulation – when and how to optimally introduce or remove it – is also a topic covered by a wide stream of literature (see (Hahn and Hird 1991; Joskow and Rose 1989; Parker and Kirkpatrick 2012) for reviews).

Empirical Literature

In this section we will review some of the empirical literature regarding the effects of government interventions undertaken to address apparent information failure in various markets.

Jin and Leslie (2003) provide one of the most thorough examinations of the results of mandatory government quality disclosure we have come across, and they claim theirs is the first to test and confirm the theory that quality disclosure leads directly to quality improvement. The authors exploit a sudden change of restaurant hygiene report cards from voluntary to mandatory disclosure. They show that the

report cards caused restaurants to increase quality, consumers increased their sensitivity to hygiene quality changes, and illnesses related to restaurant hygiene decreased in Los Angeles County. This work illustrates the fact that government information disclosure can indeed ameliorate consumer information problems. Since hygiene is generally difficult for consumers to observe, increasing availability of information about hygiene causes consumers to make more informed choices. This finding is supported in lab studies of consumer behavior as well. In one study, consumers were presented with product labels of potentially hazardous products and then asked to comment on their intended use patterns. The study found that consumers' intended use of household products (e.g., bleach and drain opener) is appropriately attenuated when the products are labeled to indicate the hazards (Viscusi et al. 1986).

The positive effects of information provision on consumer choice have extended beyond the lab as well. School choice, for example, is difficult for parents. While parents may suspect a school is under-performing, it is difficult for them to be sure enough to invest the considerable time, energy and cost of moving their child to a different school. In a natural experiment followed by a field experiment, Hastings and Weinstein (2008) examined the effects of directly providing school test scores to parents. Parents who received information about a low-performing school were more likely to move their child to a better-performing school when such a move was possible.

While the papers described above show that government intervention can have positive effects on both seller-chosen quality and consumer decision-making, there are a number of examples where interventions had little effect, or even negative or perverse effects, on consumer behavior. Several of these examples come from the healthcare services market, which, as we have discussed, is plagued with a variety of information failures (Beales et al. 1981; Dranove and Jin 2010; Dulleck and Kerschbamer 2006). A stream of literature has shown that despite efforts to provide quality information in the market, such efforts may have little or no effect on consumers' choices and may also have negative effects on seller behavior. For example, in 1999 and 2000 Medicare enrollees received report cards on HMO quality. Researchers measured the effects on consumers' subsequent choice of health plans; they found that while consumers were somewhat responsive to government-mandated report cards, market-based information dominated their health plan choices. In other words, the government intervention had little effect (Dafny and Dranove 2008). In a related analysis of hospital report cards, Dranove and Sfekas (2008) found that hospital report cards did not significantly change consumers' behavior in the market for hospital services.

In some cases, mandatory disclosure has been shown to have a negative effect when it elicits unexpected responses from consumers, as in the case of conflict of interest disclosure. Conflicts of interest represent an information failure in many services, such as medicine, law, and financial advising, which are all credence products. For example, when a doctor offers diagnosis and treatment as a bundled

service, there is an inherent conflict of interest because the diagnosis and recommendation for surgery directly profit the doctor (Gruber et al. 1999). Government-mandated conflict of interest disclosure is a purported solution to such concerns, as in the case of doctors providing informed consent (White 2004). Perverse effects of conflict of interest disclosure were observed in an experiment by Cain et al. (2005), which varied disclosure of conflict of interest. Researchers found that consumers may trust an advisor more as a result of disclosure, perhaps because disclosure is perceived to be a demonstrably honest act, even when it has been mandated. This is one example where a government mandate may cause a negative rather than a positive effect on consumer information use.

As discussed above, measurement of a small subset of quality attributes or reliance on a signal can theoretically cause sellers to shift effort to increasing their signal at the expense of their underlying or overall product quality. One instance of this is seen in the nursing home market, where consumers rarely have sufficient information about quality variations within their choice set. In some states, mandatory nursing home standards are required and disclosed to consumers. In a fascinating paper, Lu (2012) demonstrated that mandatory government disclosure of an incomplete set of quality attributes did not lead to an increase in overall quality, but rather led to a shift of effort from the undisclosed to the disclosed list of attributes. If the disclosed attributes are of less importance than the undisclosed attributes, it follows that overall quality is decreased. Some argue that this is indeed what occurs when public school accountability and quality are tied to standardized testing (Neal

and Schanzenbach 2010). Another example of a shift in effort was observed when the federal government sought to induce sellers to increase the healthfulness of their products by requiring nutrition labeling through the Nutrition Labeling and Education Act. Instead of the intended effect of increasing overall *nutrition*, the government intervention caused an overall increase in the *tastiness* of foods (Moorman et al. 2012).

These papers represent a significant body of literature that examines how government intervention can have positive, negligible, or negative effects on markets where market-based information is known to be insufficient. However, these markets are not static, and market-based information available today is often markedly different from the information available when some government interventions were established. The internet did not exist when state medical boards were established to maintain quality in doctor services. Indeed, online reviews represent a relatively new market-based mechanism of consumer information provision that has been neglected in the empirical literature of imperfect information and government intervention.

In a review paper on quality disclosure and certification, Dranove and Jin define quality disclosure as "an effort by a certification agency to systematically measure and report product quality for a nontrivial percentage of products in a market" (Dranove and Jin 2010, p. 936). Dranove and Jin surmise that online reviews occupy a blurry space between a quality disclosure mechanism and a town-square type of information source. Online reviews have only grown in popularity and volume, suggesting the time for their study in this context has arrived.

Research surrounding online reviews is widely available. Scholars have focused on the impacts of the valence, variance, and volume of reviews on prices, revenue, and sales growth. Unfortunately this research is primarily centered on search and experience products, for which market-based quality mechanisms are already the norm (i.e., there is little recognized need for government intervention). There are very few studies on consumer quality information in markets for credence products such as mechanics and healthcare services, for which government quality disclosure is often offered because no effective market-based mechanisms exist. We believe that there are two dominant reasons for the lack of empirical work in this field. First, broadly available consumer-supplied quality information in the form of online reviews is a recent phenomenon that has only become possible with the establishment of ubiquitous internet, comfort with internet transactions, and a critical mass of consumers who share their product knowledge. Second, current theory on government and third-party quality disclosure largely rests on the key assumption that in certain markets (such as those for used cars, mechanics, and healthcare services), consumers have no market-based source for quality information. Online reviews now represent a challenge to this key assumption.

More broadly, work that specifically compares an existing government regulation to a new market-based mechanism has been scarce. As discussed in the empirical section above, comparisons between market information and government-issued information are usually made by comparing existing market information to new government-issued information. The lack of a converse comparison between new

market-based quality mechanisms and well-established government mechanisms is likely explained precisely because new market-based mechanisms are not available, not feasible, too costly, or because risk to the consumer is too high (as is the case for healthcare). Alternatively, comparison may be lacking because scholars do not recognize online reviews and other emergent information sources as market mechanisms that are capable of competing with established mechanisms.

One notable exception to this paucity of research is a paper by Kang, Choi, Kuznetsova, and Luca (2013), which we briefly introduced in Chapter 1. Kang et al. create predictive models of severe restaurant hygiene violations using online review data as predictors. The authors suggest that mining online reviews may help government regulators target specific restaurants for inspection, as well as provide consumers with an alternative information source (besides government disclosure) for hygiene information. This work highlights two key departures from commonly conducted studies. First, it uses a rare event measure of quality against which to compare online reviews rather attempting to use a measure that is available for all restaurants. Rather than having a quality measure for each restaurant, as in Jin and Leslie (2003), Kang et al. use a quality measure only applied to the lowest quality restaurants. Second, perhaps as a direct consequence of the rare event outcome measure, Kang et al. follow predictive modeling methods and norms rather than the more common explanatory modeling (Shmueli and Koppius 2011) in order to investigate their question. We will discuss the appropriateness of predictive modeling for scientific inquiry in more detail below.

Our work differs from Kang et al. in both empirical setting and key feature choices. Our empirical setting is the market for doctor services rather than restaurant services. A key difference in our analysis lies in the choice of available features as predictors. Kang et al. include prior inspection results and prior post-inspection reviews as predictors, and thus the post-inspection reviews analyzed in their study may be influenced by the previous actions of the government and/or consumers. We focus on predicting unsuitability of doctors before any institutional quality disclosure has taken place. We drop all reviews that occur after a sanction, thus testing the predictive power of consumer-created reviews independent of the official government quality disclosure.

Doctors as Credence Products

As we have noted above, doctors are classic credence products. Governments require doctors to obtain medical licenses in hopes of ensuring a minimum level of knowledge and training (see Appendix, p. 109, for a sample of state medical board mission statements). When licensing authorities discover a doctor is unsuitable for medical practice, the authority will remove the doctor from practice by suspending or revoking the medical license, by requiring re-training, or by other remedial actions.

Traditionally, consumers' information about doctor quality has been limited to government licensing and word of mouth referral. In the last decade, however, consumers have turned to online reviews to share information about their doctors and research new providers. Online reviews have been found to function in a way that is

similar to traditional word of mouth, and it is now well-accepted that online word of mouth information has a significant impact on consumer decision-making for an array of products. In particular, online reviews of doctors have been studied in a number of ways. Doctor reviews have been found to correlate with population-representative patient satisfaction scores (Gao et al. 2011) and with 14-day checkup rates (Wallace et al. 2014), and thus they are useful to the consumer by measuring a doctor's ability to relate well to patients. From our research in Essay 1, we know that online reviews contain reference to doctors' knowledge, bedside manner, and a mix of other attributes.

However, it is unclear whether the content of online reviews is predictive of a doctor's suitability for medical practice, i.e., whether a doctor has the appropriate knowledge, skill, and ethics to treat patients well and fairly. Credence theory predicts that because the average patient cannot assess the appropriateness and quality of a doctor's diagnosis and treatment, the average online review will not contain that information. Despite this theory's prediction, our results from Essay 1 establish that online reviews contain claims about credence attributes. This leads to our hypothesis that *online reviews provide a signal of doctors' suitability to practice medicine*. We will empirically examine this question.

If our first hypothesis is true and online reviews can predict sanctions, this result may be driven by either credence attribute evaluations or experience attribute evaluations. If experience evaluations are correlated with suitability to practice, then experience evaluations will predict sanctions. For example, a doctor who tends to

over-treat his patients (unobservable) may also be less warm or friendly (observable). Patient satisfaction has been shown to relate to clinical outcomes both positively and negatively (Fenton et al. 2012; Luxford 2012). In Essay 1, Figure 3, p.25, we showed that approximately 75% of the attribute content of reviews is composed of experience attribute information. Given the high proportion of experience content in reviews, if in fact we do find online reviews contain a signal of suitability to practice, we expect the *correlation to be driven by observable factors (e.g., punctuality) included in the review rather than unobservable factors (e.g., knowledge).*

Medical licensing authorities may revoke a license for service failures other than defrauding patients, including observable reasons such as sexual harassment or being inebriated while at work. If there is correlation between online reviews and suitability to practice, *we expect the relationship will be driven by observable service failures.*

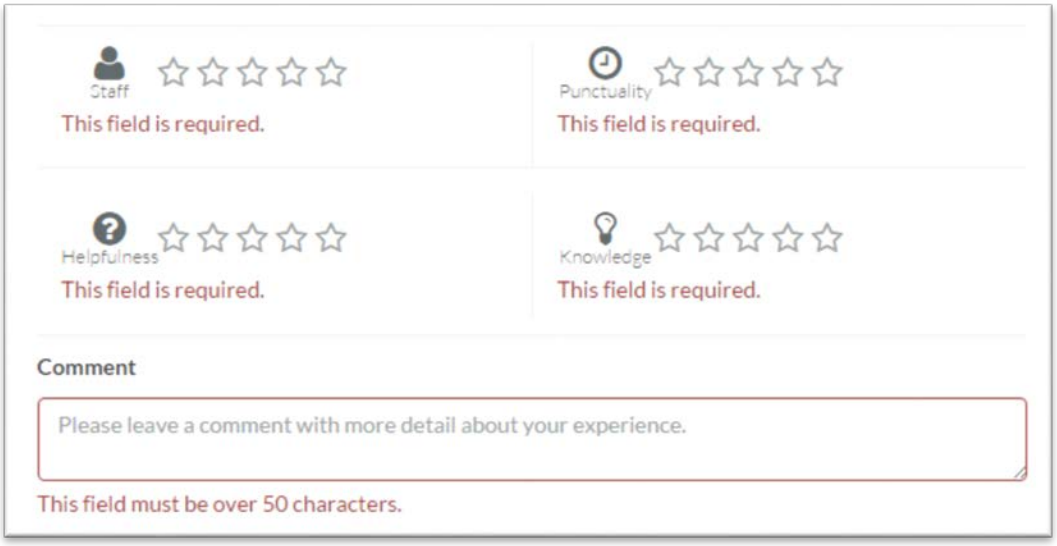
Data

The suitability of a doctor is a notoriously difficult construct to measure and obtain (Harris and Buntin 2008; Scholle et al. 2009). State medical boards establish license standards, review and rule on complaints, and make decisions on punishment terms for doctors they deem unsuitable for practice, including license suspension, revocation, or probation. State medical boards do not review every doctor; rather, they review and make rulings on cases in which a doctor's poor patient care or other behavior is particularly egregious and worthy of action.

To obtain state medical board ("board" for short) sanction data, we have collaborated with the Federation of State Medical Boards (FSMB). FSMB is a member organization for state medical licensing authorities. It collects information on all board sanctions across all US states, providing a unique opportunity for measuring doctor quality across the country. The dataset includes records for all licensed (and formerly licensed) doctors in the United States (~1.6 million doctors), basic demographic information on each doctor, doctors' license information (~700,000 have active licenses), board sanction dates, case outcomes, and explanations of the board's actions.

We link FSMB data with online consumer ratings from the largest doctor review site, RateMDs.com (Gao et al. 2012). With permission from the site owner, we gathered data on all doctors in the US with at least one rating (230,000+ records), from 2004 to 2011. RateMDs.com provides a template for consumers to leave a textual review as well as individual (1-5 star) ratings for staff, punctuality, helpfulness, and knowledge. Figure 5 shows what the user sees when leaving a rating. RateMDs.com provides a summary of doctor ratings by averaging each doctor's rating across categories and offering an "overall" quality score comprising an average of only the helpfulness and knowledge ratings. We use the physician scores in the punctuality, helpfulness, and knowledge categories, as the staff category was only introduced in 2008 and was therefore missing from about half of our data.

Figure 5: Screen capture of RateMDs.com doctor ratings input



From the FSMB data, we have the following covariates: medical school graduation year, medical specialty, state of license, zip code of doctor's home address, severity of sanction (restriction, probation, suspension or revocation), description of the basis for the sanction (e.g., overprescribing medications), and medical school.

The database linking process (i.e., matching FSMB doctor records to RateMDs.com doctor records) resulted in a sample of 141,961 doctors who have at least one online rating. This represents a subsample of the original data because we discarded any records that could not be perfectly linked between databases. Within this group, 6,840 doctors have received medical board sanctions at some point in their career. Table 8 (several sections below) presents summary statistics of the average RateMDs.com ratings for the full sample, the subsample of doctors who have never received a sanction, and the subsample of doctors who have received a sanction.

Method

Predictive Modeling

Our hypotheses are fundamentally correlational and predictive. There is no suggestion that online reviews may *cause* a doctor to be sanctioned by a medical board. Rather, our hypotheses suggest that there is a predictive signal of doctor suitability that is retrievable from online reviews prior to a medical board sanction. That is, while there may be noise in online reviews, our goal is to test whether or not there is at least some true signal that we can use to make accurate predictions about future sanctions.

This goal is in contrast to the traditional approach to positivist research that dominates the Information Systems literature (Agarwal and Dhar 2014; Shmueli and Koppius 2011; Shmueli 2010). Many papers investigating online review behavior aim to explain causal relationships with, rather than predict, product quality. For instance, some aim to explain the causal impact of reviews on buyer behavior, e.g., whether a higher online rating increases willingness to buy, independent of true quality (Chevalier and Mayzlin 2006). Others seek to understand how online reviews impact whole markets and to assess the differential impact of specific review features, e.g., variance of beer ratings or selection of restaurants (Clemons et al. 2006; Luca 2011). Still others seek to measure how online review information differs from other sources of information, e.g., patient satisfaction scores from representative population samples differ significantly from online review ratings, which is as expected due to

selection biases (Gao et al. 2011). These are excellent examples of causal, explanatory theory-testing. They differ from the focus of this research in that they aim to carefully measure and explain the existing impact of online reviews on behavior. In contrast, we aim to carefully measure and predict the potential impact of online reviews, i.e., whether consumers *should* use online reviews as a means of avoiding low-quality credence products.

We recognize this is a significant departure from tradition. While predictive papers are rare in Information Systems research, the underutilization of predictive methods is not indicative of their lack of value. Rather, when the research question calls for a predictive context, IS researchers must design their methods accordingly (Shmueli and Koppius 2011). Further, our work is in the domain of healthcare, where the value of prediction can be measured in lives saved (Agarwal and Dhar 2014; Bardhan et al. 2015).

Credence products theory proposes that consumers lack quality information in a forward-looking decision context. Our paper tests that theory by searching for the existence of quality information from online reviews in the consumers' decision context, similar to work by Kang et al. (2013). In this framework, we use predictive modeling methods that are designed for detecting signals and preventing potentially misleading results generated by pitfalls such as over-fitting, i.e., results that suggest more predictive power in a retrospective model than its use on future data would support. Our predictive model answers the question, "Are online reviews useful to avoid bad doctors?" Answering this question simultaneously tests our first

hypothesis, the prediction that online reviews contain a signal of a doctor's suitability to practice medicine.

Feature Engineering

Feature engineering refers to the process of collecting, cleaning, and transforming data into variables to be used in a predictive model. Unlike in explanatory modeling where input variables are chosen to match theoretical constructs, variables in predictive models are chosen according to how well they can predict the outcome. Of course, there is likely to be a strong correlation between variables chosen for explanatory versus predictive models, but the correlation is not perfect. For example, in predictive models all features, X, must precede the outcome, Y, in chronology. Also, we have the additional constraint that not only must X precede Y, but X must also be available at the time of prediction (Shmueli 2010). Leakage is defined as data about the outcome that is not a legitimate source of data in the real decision setting (Kaufman et al. 2012). For example, during the process of data collection, cleaning, and feature engineering, it is possible for posterior signals of the outcome variable to be mixed into the features used to predict that outcome, causing tautological prediction, e.g., "It rains on rainy days." (Kaufman et al. 2012, p. 15). In our case, we wish to investigate whether online reviews contain signals of a doctor's suitability to practice medicine, independent of prior medical board actions. To avoid leakage, we dropped all reviews for a doctor that occur the month of or the months after that doctor's medical board sanction. We chose to include reviews

leading up to the official medical board sanction because they are available to the consumer as real information prior to an official medical board sanction. We discarded observations (doctors) that receive no reviews before a sanction, bringing our sample to 134,973.

Next, we aggregated ratings for each doctor into summary features. For each doctor's rating on RateMDs.com, there are individual star ratings for punctuality, knowledge, and helpfulness. To create features at the doctor-observation level, we use the average of the knowledge, punctuality, and helpfulness ratings. We also include the count of reviews for each doctor. These numerical features (i.e. valence and volume) are often used in online review research. Variance is another metric used to describe reviews (Sun 2012), however 43% of our sample doctors only have one pre-sanction review, thus variance is null for those observations. This lack of variance would likely be strongly influenced by the outcome variable, and thus subject to leakage. Therefore we did not include variance in our models. Similarly, the number of removed reviews (purportedly by the site's spam filter), the ratings selection counts (e.g., count of ones, count of twos, etc.), and the proportions of the ratings selections (e.g., proportion of ones, proportion of twos, etc.) are likely to be influenced strongly by the outcome variables, would likely introduce leakage, and thus we did not use them.

We examined demographic variables (e.g., specialty, state of license, and graduation year) as potential features. In some cases there were less than 30 observations in a demographic class (i.e., Nuclear Medicine and Medical Genetics).

In the context of explanatory modeling, these variables might be discarded from the explanatory model because they cause bias in the beta estimates resulting from multicollinearity or because they are so rare that they almost perfectly predict the outcome (i.e., close to rank deficiency in the X matrix). While we are less concerned with bias of beta estimates in the context of prediction, these models can cause overfitting. Instead of dropping them from the model, predictive modelers use feature reduction techniques such as Principle Components Analysis (PCA), ridge regression, lasso, and elastic nets (discussed more below).

Our demographic variables included many missing observations. In the case of specialty, our sample included 2,243 doctors with no known specialty. Instead of dropping these observations or imputing the sample mean, as is common for explanatory modeling, we used missingness as a feature because it is likely to be meaningful in this context. Doctors who do not register with a specialty board may behave differently in the market, and thus we wanted to capture this potentially predictive information. We created a category "UNKNOWN" and allowed the observations to remain in the model with a binary indicator of missingness. This is acceptable for classification methods (Ding and Simonoff 2010).

Text Features. We followed a series of text mining steps to create textual features from the written content of the reviews. As with the ratings, we first rolled the textual reviews at the review-observation level up to the doctor-observation level by concatenating each doctor's textual reviews into one corpus. From each corpus, we

extracted text-based features using LightSide Researcher's Workbench, a graphic user interface for text feature extraction and analysis.

Textual data includes many irrelevant pieces of information and strings, which text miners normally discard at the outset of feature engineering in order to reduce noise in the dataset (Boyd-Graber et al. 2014; Han et al. 2016). We first reduced the text data by removing stopwords (e.g., and, or). Next, we removed words that occur less than 5 times in the data. While removing very rare words from the feature set is common, we tested whether it was warranted in our context by comparing models generated with and without this step. We found no differences other than length of time to generate the features and train the models. Thus, we left stopwords out of our feature space. Similarly, we tested and removed punctuation. In our next step, we "stemmed" the words so that "doctor" and "doctors" would count as one feature rather than two. We call the stemmed words "tokens," i.e., "doctor" is the token that represents the words "doctor" and "doctors."

After the preparation steps, we generated several document term maps. A document term map is a matrix of rows and columns where the rows represent observations (i.e., individual doctors) and the columns represent each token in the feature set. The cells contain frequencies of the tokens for a given observation. We extracted several sets of features: unigrams (one-word "tokens"); bigrams, trigrams; and part-of-speech bigrams. Unigrams are equivalent to one binary variable feature for each word. This method assumes no value from the ordering or sentence structure of the words; in other words, the relevant signal is carried in the meaning behind

individual words and not in their position in relation to one another. Unigram features alone can be powerful predictors, as shown by hygiene words in Kang et al. (2013). While this often gives positive results, it may leave some meaning, and therefore signal, behind. We generated several additional feature sets to test whether word ordering matters: bigrams, trigrams, and part-of-speech bigrams. Bigrams create one column feature for each two-word string, and similarly, trigrams contain three-word strings. Part-of-speech bigrams capture negation, such as "the doctor's bedside manner was **not good**."

In addition to the standard raw count of frequency in our document term matrices, we also tested the use of term frequency-inverse document frequency (TF-IDF) weighting. This weighting normalized the frequency of the term in a document by the frequency of the term in the overall dataset (Manning et al. 2009). In our case, this weighting did not produce additional predictive power, and thus we do not report additional results from the TF-IDF feature set.

Model Selection

Our prediction challenge is a supervised binary classification challenge. Classification refers to our goal of using our features to predict a discrete category for each observation, a.k.a. class membership. Classification algorithms can predict multiple discrete classes; however, our outcome variable only has two values, sanctioned or not sanctioned, and is therefore "binary." "Supervised" refers to the fact

that we know the true value of our outcome and can train and evaluate our models against the true values.

There is a wide range of classification algorithms available, which includes parametric and non-parametric methods. Common approaches include decision trees, support vector machines, and regression. These approaches are all suitable for supervised binary classification; model selection is guided by model performance. We trained each of these classifiers using the same ratings and text feature sets and compared area under the receiving operator curve (AUC) values for each (Hanley and McNeil 1982). In every comparison, logistic regression outperformed other algorithms, and so we report our results using logistic regression.

We use AUC as a measure of model performance, rather than accuracy, because we have an imbalance between our classes. While accuracy is an intuitive evaluation of a classifier, it can be misleading with imbalanced data. For example, a naive model that simply classifies all doctors as non-sanctioned would immediately achieve 98.98% accuracy, because only 1.02% of the doctors are sanctioned. Area under the ROC curve is more appropriate for evaluating classifiers of rare events (Fawcett 2006). The AUC is equivalent to the probability that the classifier would rank a randomly selected positive instance higher than a randomly selected negative instance. Thus, an AUC of 0.50 represents random chance, and anything above 0.50 represents the predictive power of the model.

We first trained our models on individual sets of features, e.g., first we trained a model using only the ratings data, then the text data, then combinations of those

sets. In explanatory modeling, theory drives control and explanatory variable selection. In predictive modeling, we use the features that are most effective in predicting the outcome, barring leakage. Overfitting is a concern, however. We "overfit" a model when the model's performance is higher than what it would be on a new real-world dataset. In our case, we used 10-fold cross-validation to train our models and find the best regularization constants, then used a 30% randomly selected holdout set to report final model performance.

In addition to algorithm selection, we also compared regularization options. Regularization is a method of weighting features to reduce the dimension of the feature space. This method reduces overfitting by reducing the effects of multicollinearity of features. It is not used in explanatory modeling because, like PCA, it obfuscates the meaning of each predictor variable, thereby reducing interpretability of the model and alignment to theory. We compared L_1 and L_2 regularization options empirically. The L_1 reduces many beta-weights to zero, and it is reasonable in a case where many features are completely irrelevant (Ng 2004). The L_2 scheme penalizes higher sum of the squares of parameters. Since we are investigating many textual features and specialties, we let model performance be the determining factor in our choice of regularization scheme. In every case, L_2 regularization outperformed L_1 or no regularization. Therefore, we used and report the results using L_2 regularization.

Our feature preparation and modeling were performed using R and LightSide Researchers Workbench, which is a WEKA-wrapper for text mining (Mayfield and Rosé 2013).

As discussed above, we compare model performance using area under the receiver operating characteristic curve (AUC) and inspection of the receiver operating characteristic (ROC) curve (Fawcett 2006; Hanley and McNeil 1982). The ROC curve is a visual method of evaluating the performance of a classifier at any cutoff point, plotted from the classifier's false positive rate (horizontal axis) against the true positive rate (vertical axis). It is particularly useful for evaluating classifiers with uneven class distributions or "rare" events. The ROC curve is also useful for inspecting the qualitative model performance at different error costs (which in our case might be a consumer with a higher tolerance for false positives than for false negatives, (Fawcett 2006)). The AUC is the area under the ROC curve, and it is "equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" (Fawcett 2006, p. 868). Direct comparison of AUC and ROC of different classifiers is a commonly accepted method of classifier evaluation. We further report the applied implications of our results using a lift curve and through discussion of appropriate scenario-based cutoff values.

Results

In Table 8 we present summary statistics for the online ratings across the full sample and in the sanctioned versus unsanctioned doctors. The summary statistics for the unsanctioned doctors are almost identical to the full sample because sanctioned doctors are a small subset. Sanctioned doctors have lower ratings across all categories. The review count was artificially reduced only for sanctioned doctors because we dropped reviews that occur after a sanction. Therefore we do not draw conclusions on the comparison of the review count statistics or on the length of the textual reviews.

Table 8: Summary statistics for the full sample, unsanctioned doctors, and sanctioned doctors

	Full sample (N=134,973)		Unsanctioned doctors (n = 133,600)		Sanctioned doctors (n = 1,373)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Number of Reviews	2.99	3.39	2.99	3.39	2.97^	3.59
Avg. Knowledge	4.03	1.25	4.04	1.25	3.73	1.37
Avg. Helpful	3.89	1.36	3.89	1.36	3.57	1.47
Avg. Punctual	3.83	1.18	3.83	1.18	3.42	1.31

^Note that since we drop sanctioned doctors' post-sanction reviews, the mean number of reviews for unsanctioned versus sanctioned doctors cannot be directly compared.

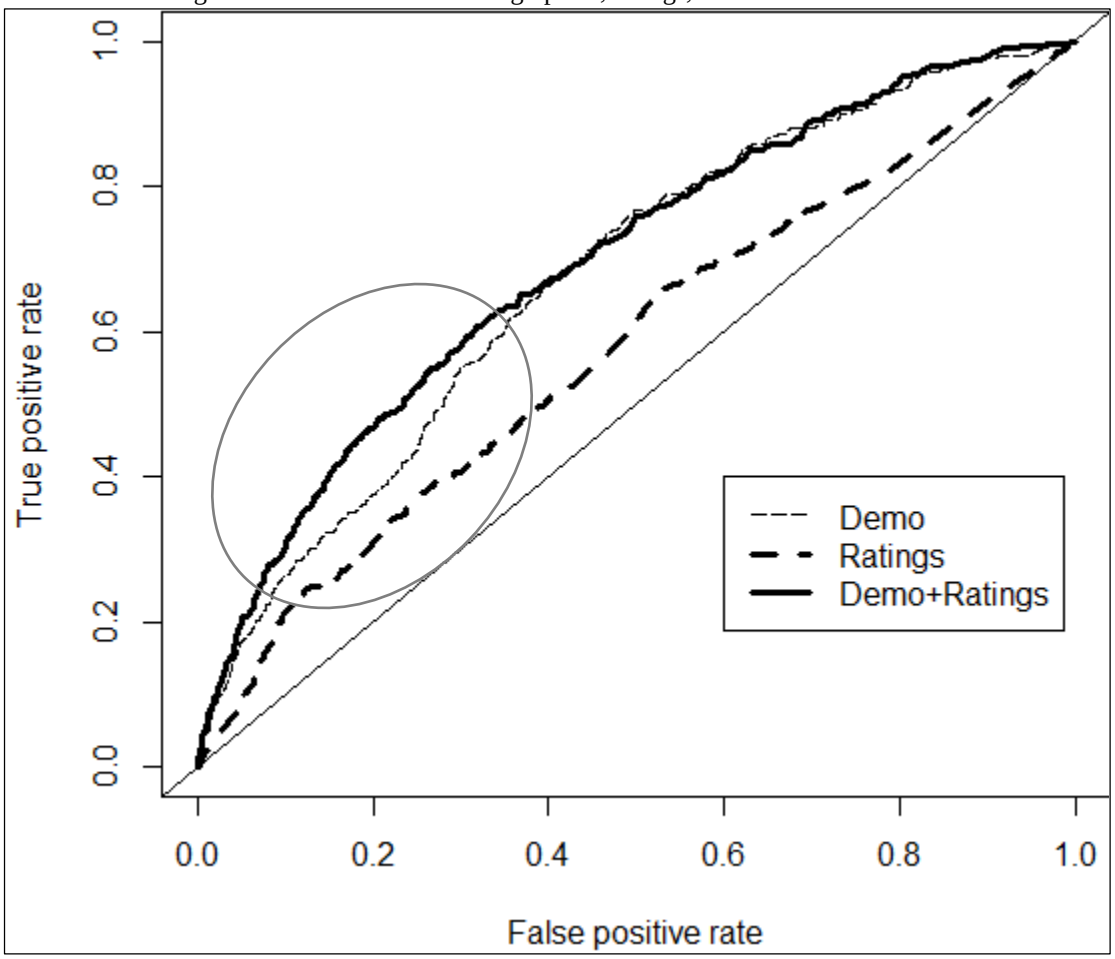
To determine whether the differences in ratings can be useful in distinguishing a doctor's suitability to practice ex ante, we trained a L_2-regularized logit classifier using demographics and ratings variables. We find that the combination of online ratings of doctors and their demographic information has significant predictive power (AUC= 0.695). The predictive power of our models is largely driven by the demographic information (AUC=0.675). RateMDs.com ratings variables by

themselves, inclusive of the number of reviews, average knowledge rating, average helpfulness, and average punctuality resulted in an AUC of 0.576. While small, the increase in AUC as we shift from demographic information only to the combination of demographic information and online ratings belies the independent and practically relevant predictive power of the online ratings (Pencina et al. 2008). Area under the ROC curve results for our initial models are given in Table 9. The ROC curves are given in Figure 6.

Table 9: Area under the curve (AUC) for logit models

Model Variables	**AUC**
Demographics	0.675
Ratings (review count, knowledge, punctuality, helpfulness)	0.576
Demo + Ratings Combined	0.695

Figure 6: ROC curves for demographics, ratings, and combined models



Note: The circled region indicates the range of classifier cutoffs where the ratings provide add predictive power over to the demographic variables.

An examination of the ROC curves indicates that online ratings only add predictive power to the demographic information in the lower left region of the graph. The lower left region represents more conservative model cutoff points, i.e., fewer false positives but also fewer true positives. We will discuss practical implications of decision rules for choosing a doctor, cutoff selection (i.e. which region of the graph to consider), and performance in the general discussion section. First, we examine the

role of unstructured text in predicting a doctor's suitability to practice medicine in the section below.

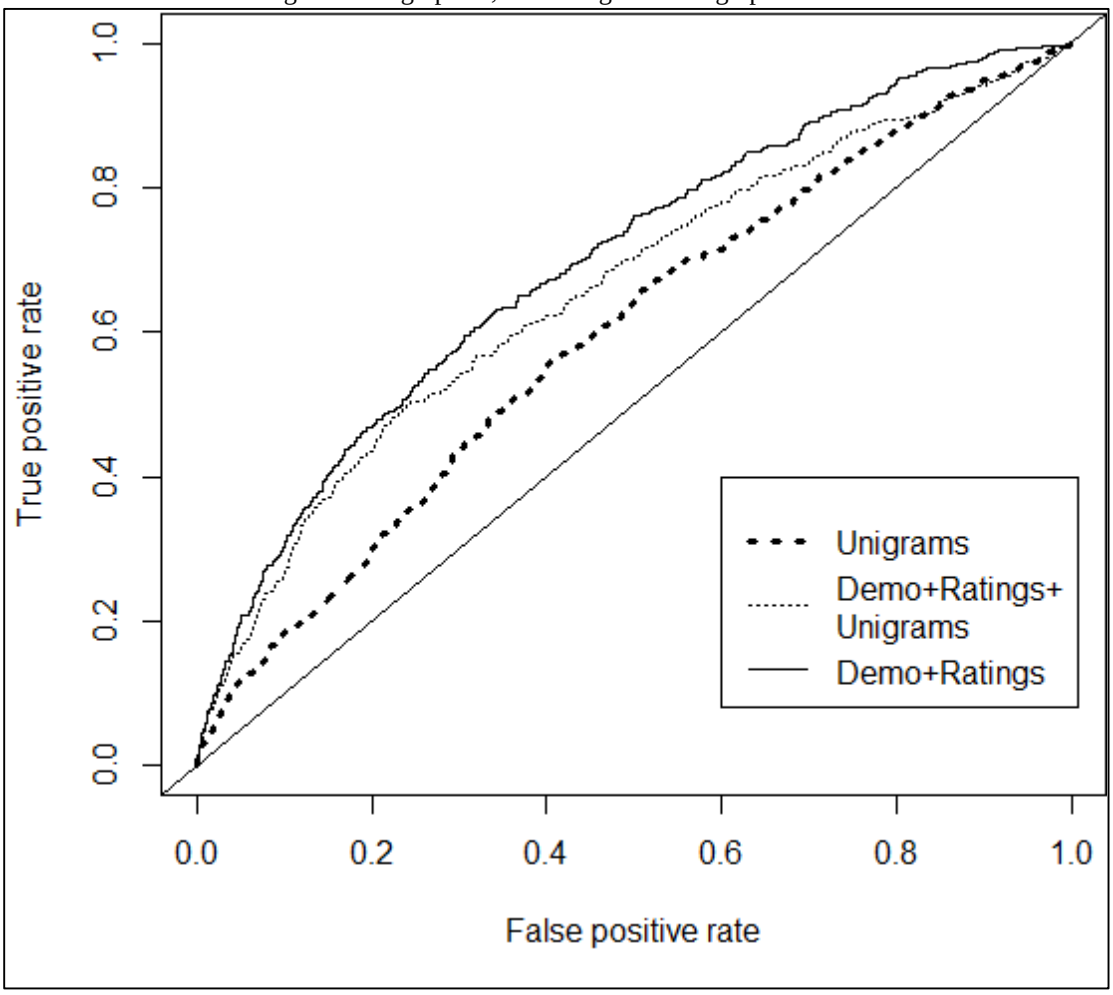
Text Features of Reviews

Results of our text-only models are reported in Table 10 and Figure 7. The text feature sets provided predictive power in a model by themselves (i.e., AUC ~ 0.60 for both unigrams and bigrams). However, no text feature set increased the AUC over the models reported above, which used only demographics and ratings as predictors. The ROC curves further demonstrate that there is no *added* predictive power from text features over and above the ratings and demographics features. At all points, the combination of demographics and ratings alone outperforms the text models, which is shown in Figure 7, where the demographics+ratings ROC curve is to the "northwest" of the other curves. To maintain clarity, Figure 7 only shows the ROC comparison for unigram features, commonly known as "Bag of Words." However all text feature sets (unigrams, bigrams, and part of speech bigrams) were the same in that they did not increase AUC over the model of ratings and demographics alone on any area of the ROC curve. The lack of additional predictive power of unigrams and bigrams was a surprise to us; we discuss and explore this curious result in more detail in the discussion.

Table 10: Regularized models with text features

Model	AUC
Unigrams ("Bag of Words")	0.5956322
Bigrams	0.5955046
Part of Speech Bigrams	0.5283601
Demo + Ratings + Unigrams	0.657903

Figure 7: Comparison of ROC curves for models with text features (unigrams) alone, text together with ratings & demographics, and ratings & demographics alone



Discussion. We hypothesized that *online reviews provide a signal of doctors' suitability to practice medicine*. Our results support this hypothesis. We found that online ratings and online review text both carry information about a doctor's suitability to practice. Text, in the form of unigrams and bigrams, carries signal on its own, but is dominated by the signal from doctor demographics and ratings.

It is important to note that while demographics (i.e., specialty, state of practice, and year of graduation) are helpful in tuning the predictive model and have higher predictive power than ratings and text, they are less helpful to a consumer making a decision. A prospective consumer is likely to search for a doctor of one specialty, in one state. Absent of ratings data or year of graduation, our model would produce the same prediction for all obstetricians in Arizona, for example, and therefore not be of use to a consumer looking for obstetricians in Arizona. So, while the demographic features are important to include in the most useful model, they are not helpful without the added information from ratings.

Our results involving the text features stand in contrast to findings from Kang et al. (2013), where text features (unigrams and bigrams) provided the strongest signal of regulatory actions. It is possible that in our dataset the variation in ratings sufficiently summarizes all the variation that is available in the text. It is also possible that in Kang et al. (2013) text features are driven by prior government quality disclosures. Still, it is a puzzle why a summary numeric rating such as average punctuality would provide more predictive power than textual content.

One possible explanation is that consumers constrain their textual descriptions to the attributes listed in the ratings. On the RateMDs.com input screen, review writers are asked to first rate punctuality, knowledge, and helpfulness, then write their review. Survey research scholars have long held that the order of questions influences the results, and recommend that more general questions should precede more specific questions to avoid drawing focus to solely the more specific questions (Strack 1992).

Since RateMDs.com reviews are written *after* ratings are selected, reviewers may align their review language to support the selected numerical rating.

To explore the question of how our text features related to the ratings, we investigated the correlation between individual unigram tokens and each type of numerical rating. We compared the top twenty tokens most negatively correlated with each numerical rating (i.e., helpfulness, knowledge, and punctuality). We expected to find that most of the tokens that most negatively correlated with punctuality, for example, would for topically relevant to a doctor's punctuality (i.e., "late," "time," "appointment"), and similarly relevant words for knowledge and helpfulness. Instead, we found that the top collection of twenty words for each of the three ratings overlapped significantly, without any immediately obvious, distinct topics. Instead of 60 tokens grouped by ratings, this comparison resulted in a total of 29 unique tokens. This overlap across all three ratings categories suggests that there may be very little variation in the tokens used to describe doctors in online reviews. Table 12 presents correlations of the top 20 terms for each rating type. The table has been shaded to show how closely the correlations track with each other (i.e., darker shading is a stronger correlation with the rating). "Rude" is the most negatively correlated token with all three ratings categories, despite it being seemingly unrelated to a doctor's knowledge. The word "know" is not in this list, suggesting that negative knowledge ratings are not paired with terms specific to knowledge. There is a similar lack of topic-related tokens for helpfulness. There is not as much variation across the ratings correlations as we expected. Our intuition suggested that negative punctuality ratings

would be most strongly associated with late words (e.g. “late,” “minutes,” “wait”), and negative knowledge ratings with different words (e.g., “know,” “diagnosis”), but that is not what we found. While more investigation is warranted, this analysis suggests that textual tokens provide little variation in comparison to negative ratings variables, which we expect to have the most signal for doctor sanctions. This might explain why the tokens did not add additional predictive power to our models.

Table 11: Correlations between top unigram tokens and ratings

Correlations	Helpfulness	Knowledge	Punctuality
rude	-0.27	-0.24	-0.24
told	-0.25	-0.23	-0.19
did	-0.25	-0.22	-0.18
thi[s]	-0.25	-0.23	-0.20
wa[s]	-0.22	-0.19	-0.15
n't	-0.19	-0.17	-0.16
worst	-0.19	-0.18	-0.17
do	-0.19	-0.17	-0.17
then	-0.19	-0.17	-0.17
<SINGLEQUOTE>	-0.18	-0.16	-0.13
refus[e]	-0.17	-0.16	-0.13
unprofession[al]	-0.17	-0.16	-0.16
after	-0.16	-0.14	-0.14
anoth[er]	-0.16	-0.14	-0.14
ask	-0.16	-0.13	-0.12
back	-0.16	-0.14	-0.14
poor	-0.16	-0.14	-0.13
pain	-0.16	-0.14	-0.11
went	-0.16	-0.14	-0.12
monei	-0.15	-0.14	-0.13
out	-0.15	-0.14	-0.13
because[e]	-0.15	-0.13	-0.15
get	-0.14	-0.11	-0.15
room	-0.14	-0.12	-0.17
hour	-0.13	-0.10	-0.23
minut[e]	-0.13	-0.11	-0.16
appoint[ment]	-0.11	-0.09	-0.16
wait	-0.11	-0.08	-0.25
late	-0.08	-0.06	-0.16

Note: Unigram tokens are stemmed, meaning they are reduced to their word root. Brackets indicate the likely end of the stemmed root. For example, the "minut" represents appearance of both "minute" and "minutes" in the text corpus. The cells are shaded to indicate the strength of correlation. Stronger negative correlations are darker.

It is important to note that the unigram tokens do not fully represent all meaning within the online reviews. In Essay 1, we found that a claim with supporting evidence is most credible and most likely to increase a consumers' willingness to buy. In this essay, we are not able to identify argument quality or other nuances of written

text that are not captured by unigrams. Thus, there may be other features within the textual reviews that predict sanctions, but which we cannot extract.

Our next hypothesis predicted that any correlation between online ratings and sanctions would be driven by observable factors (e.g., punctuality) included in the review rather than by unobservable factors (e.g., knowledge). We test this hypothesis in the following section.

Observable versus unobservable attributes

We hypothesized that *correlation [between ratings and sanctions] is driven by observable factors (e.g., punctuality) included in the review rather than unobservable factors (e.g., knowledge).* RateMDs.com provides four categories of ratings: knowledge, helpfulness, punctuality, and staff. Figure 5, on page 66, showed the input screen for these four ratings. As previously mentioned, staff ratings were introduced in 2008, more than 4 years after our data begins, so we have not included them in our analysis.

To test our hypothesis, we examined the AUCs produced by individual variables in the model. We thereby examined the relative predictive power of average punctuality, average knowledge, and average helpfulness ratings.

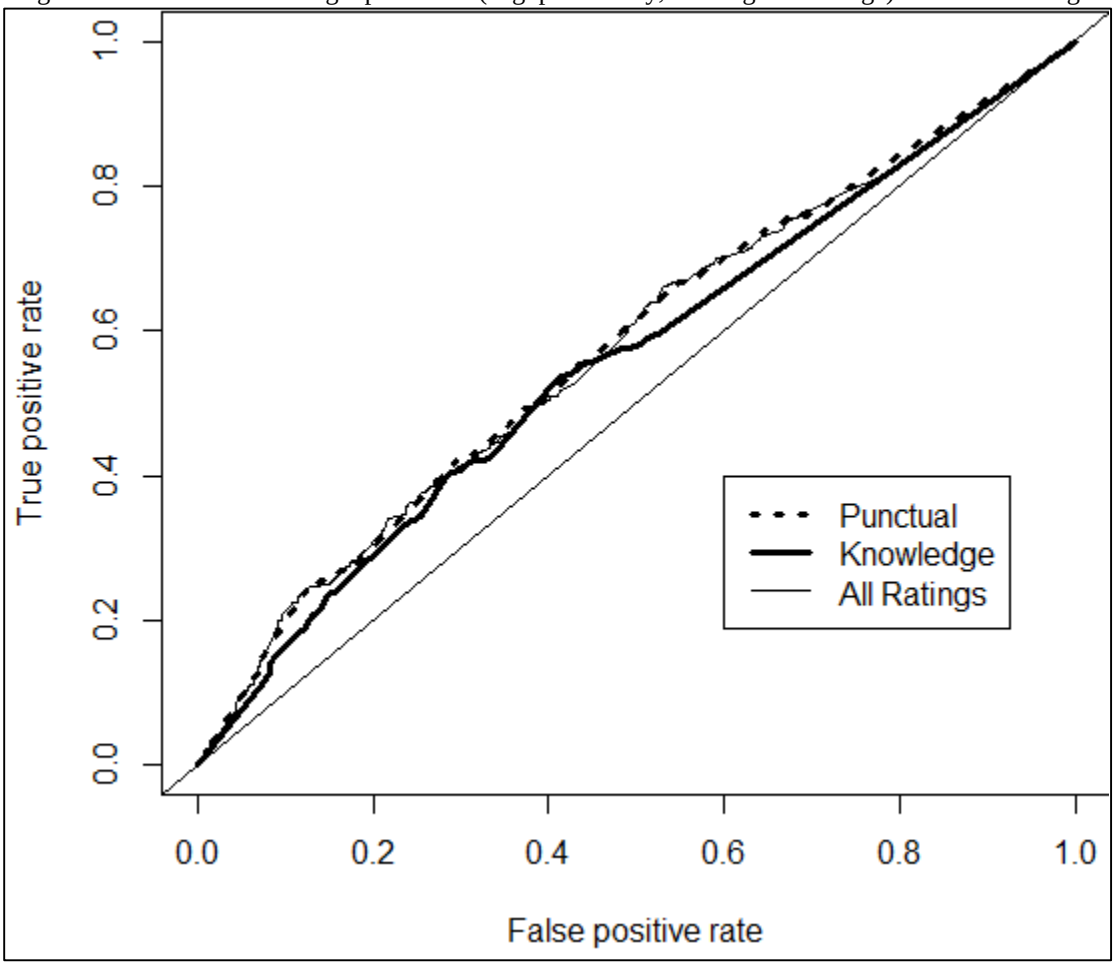
We find that average knowledge (AUC=0.56) and helpfulness (AUC=0.55) ratings are less predictive of suitability to practice than average punctuality ratings (AUC=0.58). Results for each of the online ratings variables are reported in Table 11. The ROC curves for knowledge, punctuality, and for all ratings combined are shown

in Figure 8. (The ROC curves for average knowledge and average helpfulness are almost identical, so for the sake of clarity, we do not show the latter.) In keeping with the AUC results, an examination of the ROC curves for the individual ratings also indicates that punctuality dominates knowledge as a predictor of suitability in almost all regions of the curve. A close inspection reveals that the ROC curve for all ratings closely traces the curve for punctuality; in other words, almost all of the model's predictive power is driven by punctuality.

Table 12: AUC for individual online ratings

Model	AUC
All Ratings variables (review count, knowledge, helpfulness, punctuality)	0.5761223
Review Count	0.5363439
Avg. Helpfulness	0.5534299
Avg. Knowledge	0.5596614
Avg. Punctuality	0.5784759

Figure 8: ROC curves for single predictors (avg. punctuality, and avg. knowledge) and for all ratings



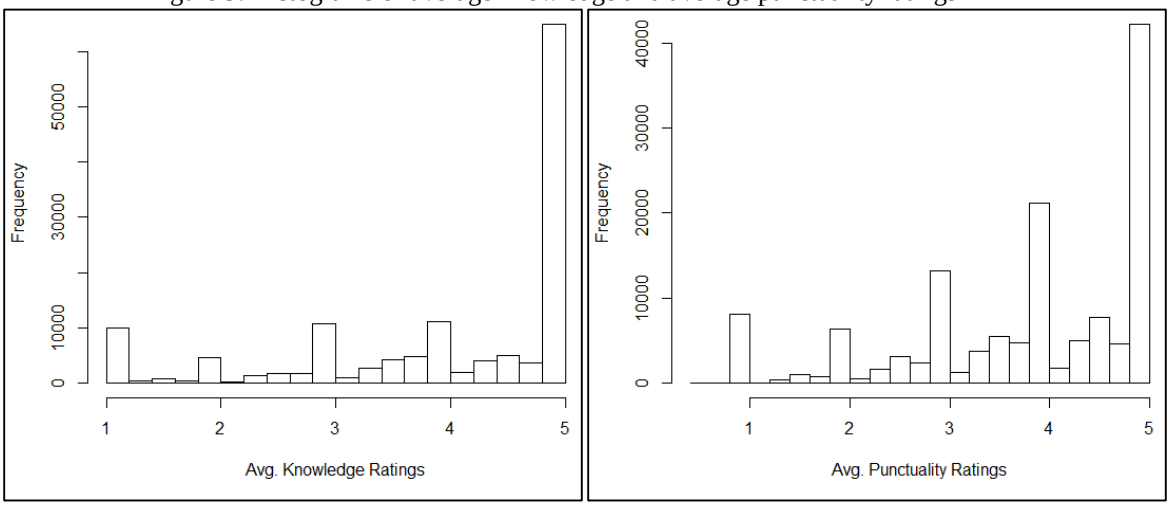
Note: Average helpfulness traces knowledge almost exactly, and thus for the sake of visual clarity we do not report the ROC curve for helpfulness.

Discussion. We hypothesized that directly observable factors would drive correlation between online ratings and a doctor's suitability to practice. This hypothesis is partially supported. We find that knowledge and helpfulness do provide some independent predictive power. We also find an association between knowledge and suitability, which, albeit small, offers evidence that a patient's perception of a

doctor's medical expertise does correlate with the doctor's actual expertise. This finding runs counter to credence theory.

Of all the ratings variables, however, punctuality provides the strongest predictive power. This association between punctuality and suitability could indicate that bad doctors are more often late. Alternatively, it could indicate that patients may have a negative feeling about a doctor and feel more confident negatively rating the doctor's punctuality than their knowledge or helpfulness. Punctuality ratings are more evenly distributed across 1s, 2s, 3s, and 4s than knowledge ratings (see Figure 9), indicating that patients observe more variation in punctuality than knowledge, which is consistent with the theory of experience and credence attributes.

Figure 9: Histograms of average knowledge and average punctuality ratings



In practice, the predictive performance difference between punctuality and the other ratings variables may have real consequences for patient choice. A patient may read reviews of her prospective choice of doctor and decide to ignore punctuality in favor of only considering knowledge and helpfulness ratings. In fact, as we will

discuss more in depth in our conclusion, this is exactly what RateMDs.com recommended during the early years of its existence. If her goal is to avoid low-suitability doctors, however, our results suggest that she should do the exact opposite: she should rely the punctuality ratings alone. We will further discuss the implications of online ratings for patient decision-making in the general discussion section. First we examine differences in observable service failures in the section below.

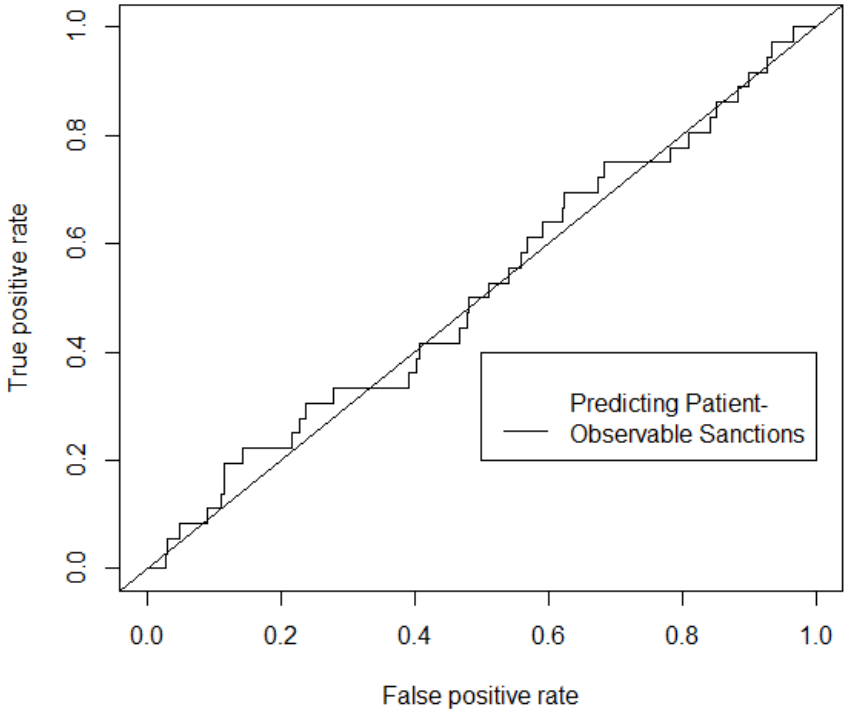
Observable service failures

We predicted that if online ratings signal a doctor's low suitability to practice medicine, the signal would pick up observable failures rather than unobservable failures. To examine the predictive nature of online ratings for observable versus unobservable suitability failures, we evaluated the performance of our models in predicting *subcategories* of medical board sanctions according to whether or not the cause for sanction was strictly observable to individual patients. The cause for sanction was identified in the FSMB data in the form of 220 distinct "bases for action." These bases ranged from generic causes such as "Gross Negligence" to "Failure to Comply with Continuing Medical Education" and "Tax Fraud." Some of the bases seemed more likely to be strictly observable to a patient (e.g., "Wrong Site Surgery") than others (e.g., "Medicare Fraud"). Two judges coded each basis as either observable or unobservable by a patient. We defined observable to be true only if at least one patient could clearly observe the behavior directly. This criteria resulted in an unacceptable interrater reliability (Cohen's kappa = 0.42 (Boudreau et al. 2001)).

We compared results from our two judges and added the criteria that a basis must be observable *in the clinical setting,* i.e., the patient must be able to observe the behavior in the course of their care and have the ability to judge the appropriateness of the action. We wrote these rules and then obtained classifications from a new independent expert judge, a pulmonology/critical care doctor. Her ratings marginally agreed with our reconciled classification (kappa = 0.62). We discussed differences between the three judges. Two additional changes were made, but no new judgment rules were created. A table of all bases for action and their "observable" codes is available in the Appendix.

Our hypothesis stated that "*we expect the relationship [between online ratings and sanctions] will be driven by observable service failures.*" We used our strongest model (i.e., demographics plus ratings) to predict observable sanctions, with observable classified as described above. The resulting model had an AUC of 0.513. The ROC curve (Figure 10) alternates above and below the chance line. Our predictive models for overall low suitability had little to no ability to predict the subset of patient-observable low quality, according to our classification of patient-observable. Our third hypothesis, therefore, is not supported.

Figure 10: ROC curve for prediction of patient-observable sanction bases



General Discussion

Our task was inherently a predictive modeling task; we sought to determine whether online ratings provide a signal of suitability to practice medicine, and thus whether they may reasonably inform a patient's prospective choice of a doctor. Economic theory suggests otherwise, positing that online reviews should offer zero information about a doctor's suitability to practice, which is consistent with the classification of healthcare services as a credence product. Our findings are contrary to this prediction: we have demonstrated that the online ratings do contain a signal, albeit low, of a doctor's suitability to practice medicine as indicated by state medical

boards, and we offer a more nuanced perspective on what online ratings can offer consumers.

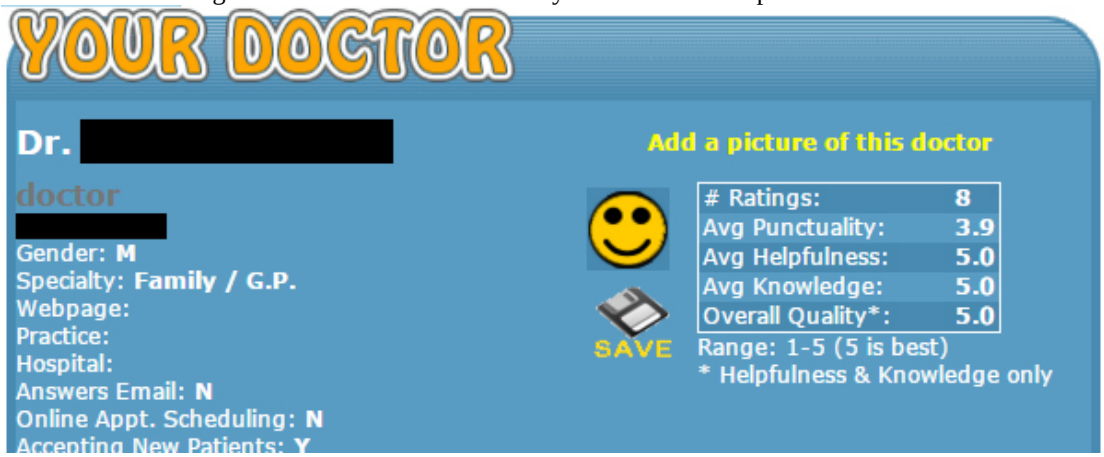
We find that directly observable attributes, such as a doctor's punctuality, correlate with suitability more than patient-perceived knowledge or helpfulness. A lower punctuality score correlates with a higher likelihood of sanctions. There are several possible interpretations of this finding, including the simplest explanation: being late to appointments correlates to poor care, which leads to state medical board sanctions. It is also possible that patients may feel more confident evaluating punctuality and therefore provide more variation in punctuality scores than in scores for other ratings variables. (Punctuality scores had the lowest mean and the highest entropy of the three individual ratings variables in our sample.)

The predictive value of punctuality may be counterintuitive to the casual observer, who may assume that out of three scores – punctuality, knowledge and helpfulness – punctuality would be the least predictive of a medical board sanction. In fact, during the time period in which our data was collected, the RateMDs.com platform purposefully omitted punctuality as a factor in its "Overall Quality" rating. Rather, RateMDs.com calculated doctors' "Overall Quality" using solely helpfulness and knowledge ratings. For example, in January 2012 Dr. Tardy[3] had a total of eight ratings. Dr. Tardy received all 5s for knowledge and helpfulness but mixed ratings for punctuality. Using only knowledge and helpfulness, RateMDs.com calculated Dr.

[3] Real doctor and screenshot, but not the doctor's real name.

Tardy's "Overall Quality" to be a 5. Figure 11 shows what the consumer would have seen on the RateMDs.com site for Dr. Tardy in 2012. We note that the design of this consumer information platform steers consumers away from the information that our study finds to be the most useful for avoiding the worst doctors. RateMDs.com may have intended to use the "Overall Average" as a nudge (Thaler and Sunstein 2009). Our research demonstrates that it was a nudge in the wrong direction.

Figure 11: Screenshot of Dr. Tardy's[3] RateMDs.com profile in 2012



We found no association between our model performance and the observability of the cause for a doctor's removal from the market. We acknowledge that our measure of service failure observability was noisy and imperfect, and this could have affected our results. We hope that better, more fine-grained measures of service failure observability will arise in new datasets in order to aid in a more thorough exploration of this question. In a separate project, we intend to examine the topics discussed in post-sanction online reviews, which may provide more insight into the nature of individual doctors' sanctions.

Overall, our work demonstrates the importance of reconsidering old economic theories of information in light of the increasing availability of information and the diversity of sources for online reviews. In this age of the internet, the wisdom of the crowd seems to be increasingly impactful and available, even in markets for credence products (Larrick and Soll 2006).

Implications

Our findings have important practical implications for review platform owners, consumers, and regulators. In place of its "Overall Ratings" score, which is an average of knowledge and helpfulness ratings, RateMDs.com could use our model to develop a classification score for each doctor based on his attributes and online reviews. While the ratings provide some predictive power, the most impactful predictive model takes demographics and location into account, weigh the input features, and provide interpretation of cutoff points. Patients would also benefit from understanding how the scores should be interpreted and what kind of performance to expect. Review platforms could also integrate suggested model cutoffs, resulting confusion matrices, and lift, which would make them valuable tools for informing customer choice.

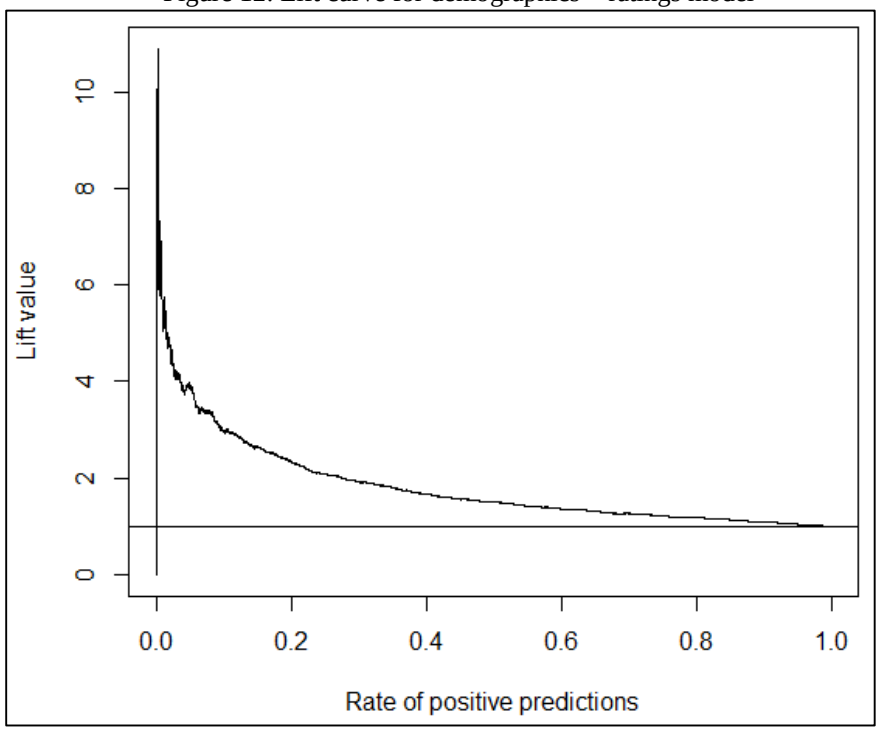
For example, in our best model, which combines demographic information and online ratings, we can choose a cutoff of 0.018 to successfully reject approximately 32% of the to-be-sanctioned doctors while mistakenly rejecting approximately 11% of doctors who haven't been sanctioned (the resulting confusion

matrix is presented in Table 12). This cutoff represents a lift of around 3 over random base rate classification. (Lift is an alternative method of representing the tradeoff between positive predictions and "hits;" see Figure 12.) A lift of 3 shows that with this cutoff, a consumer can reduce her risk of choosing a bad doctor by 1/3, provided that she is willing to give up 10% of the unsanctioned doctors from her choice set. The appropriate cutoff would be determined by individual decision criteria, such as the seriousness of a condition (e.g., a mild illness versus a life-threatening rare disease) or the potential impact of poor treatment (e.g., over-prescribing antibiotics versus a botched brain surgery). Finding a doctor for a one-time visit for a sprained ankle may be less important than finding a great obstetrician, and thus for a patient who requires the latter, it is worth sacrificing more potentially good doctors to increase confidence in rejecting bad ones (i.e., higher tolerance for false positives in order to avoid more true positives). Educating the consumer about these cutoffs so that they may make their own benefit-risk evaluations increases consumer power and information use in the market, and therefore such education promotes consumer welfare. This suggestion is in line with modern drug regulation efforts: for instance, the Food and Drug Administration is actively soliciting methods of eliciting and providing benefit-risk information so consumers can make their own informed benefit-risk tradeoffs (Food and Drug Administration 2013).

Table 13: Confusion matrix for demographics + ratings model at a 0.018 cutoff suitable for patients with a tolerance for false positives

		Predicted	
		Not Sanctioned	Sanctioned
Actual	Not Sanctioned	35731	4204
	Sanctioned	267	125

Figure 12: Lift curve for demographics + ratings model



In addition to adding value for consumer choice, our model could be used as a regulatory decision-making tool. A predictive model of doctor sanctions based on online ratings and demographics could serve as a risk-based active surveillance system for state medical boards. State medical boards currently react to complaints filed against doctors, but the Federation of State Medical Boards or individual boards could potentially take a more proactive approach and utilize active surveillance to identify egregious offenders through monitoring of online reviews. In this case, a much higher cutoff would be recommended in order to preserve resources and only "flag" deeply concerning cases (i.e., a board would select a model cutoff with a very low false positive rate). In fact, there is evidence that government regulators are already moving in this direction, turning from traditional passive surveillance

methods to prospective surveillance methods as new forms of information appear in traditionally-regulated markets. Two examples in healthcare come from the Food and Drug Administration. The first example is GenomeTrakr, a very recently established prospective surveillance method for detecting foodborne pathogen outbreaks using genomic sequencing. This strategy promises to save many lives and potentially millions of dollars by averting foodborne illness (Allard et al. 2015). The second example is the Sentinel Initiative, a prospective surveillance tool that uses electronic medical records to detect adverse events during the post-market approval phase of drug introduction (Robb et al. 2012). Outside of the healthcare industry, quality signals from online reviews are already being used in markets where regulation was typically the only method of credence quality assessment. This is seen in the market for taxis and in the transportation network Uber's use and aggressive policing of negative driver and rider reviews.[4]

We believe that the growing availability of information in traditionally information-asymmetric markets, which sometimes comes from unexpected and traditionally barren sources, will ultimately transform many credence products into experience products. The availability of information and concurrent development of methods to measure, interpret, and use the information sources will fundamentally change consumer landscapes, and therefore consumer information theory, in the coming years and decades. As online reviews grow in quantity across new product

[4] See https://newsroom.uber.com/feedback-is-a-2-way-street/ for Uber policies

categories, we expect existing signals of quality to strengthen and new ones to emerge. However, the only way to verify this is to keep testing predictive models on previously unseen data.

Limitations

Quality is a complex, multidimensional construct, and measurements of quality are subject to considerable debate (Scholle et al. 2009). We use judgements made by state medical boards as a proxy for particularly low quality, namely, a doctor's unsuitability to practice in the state. While medical board actions do not represent a comprehensive measure of quality, we suggest that revocation of a doctor's license, or probation, remediation, or reprimand by a governing body of medical professionals, is likely to be correlated with low quality such that any patient would prefer to avoid that doctor.

We also acknowledge that the measure of suitability may be conservative. False positives (i.e., cases where our model predicts a sanction where there is none) may not truly be false: the doctor's suitability may be low but not yet bad enough to garner the attention of or action by the medical board. The possibility of this scenario is supported by anecdotal evidence suggesting that strong organizational norms may prevent nurses and doctors from reporting colleagues to their medical boards. In this sense, our classifier cutoff selection and evaluation may also be conservative. That is, some of the doctors that we classify as the lowest-quality practitioners may be unsuitable to practice medicine but may also never be sanctioned by a medical board.

Conclusion

It is possible to use online ratings data to predict a doctor's suitability to practice medicine, and this finding has implications for the development of usable tools for patients who are selecting a doctor. Economic theory suggests that doctors are credence products, but our findings add nuance to that theory. Our method, predictive modeling, is suitable for testing the degree of relevance between a theory and the empirical world. Further, "predictive modeling enables assessing the distance between theory and practice, thereby serving as a 'reality check' to the relevance of theories" (Shmueli 2010, p. 4). We show that there is distance between reality and the economic predictions of credence markets, and we suspect that this distance is growing. Online ratings can help consumers, even in markets that are difficult for non-experts to evaluate, and consumers should be armed with all the predictive power our modern tools have to offer.

CHAPTER 4: CONCLUSION

Economics of information theory for credence products assumes that consumers have no access to quality information. In practice, online reviews present a challenge to that assumption. This dissertation examines the content of online reviews for credence services from a number of perspectives. We have: inspected and compared content of credence vs. experience reviews; discovered how review arguments are structured; studied consumer perceptions of the information available within online reviews for credence versus experience services; measured and discussed consumers' perceptions of the usefulness and credibility of information inside reviews; measured and discussed consumers' willingness to choose particular providers based on review content; measured and discussed the utility of reviews for helping consumers avoid low-quality service providers; and critically assessed the ability of online review metrics and content to substitute for or predict government-supplied quality information. This work contributes to theory and practice in a number of areas, as described below.

Online reviews

In Essay 1, we read reviews, systematically identified and classified the content types within, and analyzed their impact on consumer perceptions. For the services we examined, reviews were dominated by experience attribute information and comprised a mix of evidence and claims. Credence attributes were mentioned more often in credence reviews than in experience reviews. When evidence was

presented, it was more often about experience attributes. Prospective consumers of credence services found credence attribute claims to be simultaneously less credible and more important than other claims, but they were most willing to act on the claims when supported by evidence. The results highlight the impact of the review content itself, over and above the star ratings, and they illustrate the possible pitfalls of relying solely on metrics and algorithms to study online word of mouth. To the best of our knowledge, this is the first research use of in-depth content analysis methodology for extracting attribute information in online reviews.

It is difficult to compare online reviews to "true" quality because measures of true quality for reviewed products rarely exist. In the case of credence services, there are usually no effective measures of true quality for all of the sellers in the market (Dranove and Jin 2010). By linking reviews to medical board sanctions in Essay 2, this work represents one of the rare comparisons of product reviews to an alternative, objective measure of quality. We found that online reviews carry a true signal of low quality.

Economics of Information

To the best of our knowledge, we have undertaken the first research to focus on the comparison of reviews for multiple credence service providers. Given the opportunity for fraud in such services and the increasing prevalence of online reviews of credence providers, it is important to understand whether and how such reviews may be used by consumers to attempt to avoid fraudulent or low-quality providers.

Essay 1 demonstrated that credence and experience reviews contain fundamentally different information, with credence reviews including more low-credibility claims and less evidence. In Essay 2, we found that ratings can help consumers to avoid the lowest-quality doctors. This work is evidence that credence service provider quality may be more easily ascertained than has previously been the case. Economic theory relying on the assumption of zero consumer knowledge of credence service quality will need to be adjusted to account for the growing availability of online reviews. This has a direct impact on the external validity of economic lab experiments, where consumers have zero quality information (e.g., Dulleck et al. 2011). While investigating boundary conditions such as this are often useful, our work shows that quality information is likely not zero, and therefore the boundary is actually approaching zero rather than zero itself. Thus results of experiments that are designed with zero information may not be informative.

Consumer Information Platforms

Consumer information-sharing platforms exhibit cross-side network effects where more written reviews lead to a larger audience, and a larger audience leads to more reviewers (Parker and Van Alstyne 2005). To grow their network, platform owners have a strong incentive to elicit the most valuable reviews possible (i.e., high-quality, informative, and credible reviews).

In Essay 1, we test a number of factors that affect the credibility of a review. Our findings can be used by platform-makers to obtain the most credible and useful

reviews from their reviewers. Furthermore, this work highlights crucial differences in reviews for credence versus experience service providers, signaling that platform-makers should be especially careful in designing their sites to elicit only credible and hopefully important information within reviews. There is some evidence that platform owners are attempting to do this already. For example, Yahoo! encourages its users to share their "opinion" but also to be objective and to present "facts,"[5] and Yelp encourages "facts and details" in its FAQ[6] and offers "useful" in its voting buttons alongside "funny" and "cool."

RateMDs.com explicitly elicits ratings according to four attributes: punctuality, knowledge, helpfulness, and staff quality. This work demonstrates that knowledge ratings are less credible than helpfulness ratings (Essay 1), and punctuality is the most predictive of low quality (Essay 2). However, the review platform does not include punctuality when generating the overall quality rating assigned to each doctor (see Figure 11 on page 94). Consumers may not understand that this overall quality rating may not actually be very credible (Essay 1) or useful in avoiding bad doctors (Essay 2).

The credibility of online reviews has implications for consumer choice: if consumers assume the information is heterogeneous and of low credibility, they may undervalue the ratings in their decision-making, and if they assume the reverse, they may overvalue the ratings. With new reviews being posted daily, platforms should

[5]http://local.yahoo.com/review_guidelines
[6]http://www.yelp.com/faq#what_to_review

consider implementing decision models, or carefully designing choice architectures, and comparing their models against independent measures wherever possible.

Government Regulation

The new market structures of the sharing economy, e.g., Uber and AirBnB, are falling under the scrutiny of their respective jurisdictions; whether and how to regulate them is a present and pressing issue that should be informed by knowledge of the information in the market. This work may inform regulators of the potential of reviews to supply relevant information in markets where information failure is the norm and government intervention has long been accepted as the only viable response. For example, when considering regulation in a market where review volume is high, online reviews may present a viable market-based alternative or addition to government intervention.

This may be a strong claim to make with regard to the healthcare industry. However, since there is some predictive value in reviews, reviews may be useful analytical tools to evaluate and perhaps to inform regulatory decisions. Even though the industry has been highly regulated for decades, new regulations are continually developed. For example, approximately half of the states in the US do not currently license Certified Professional Midwives (CPMs) to practice their profession, and

consumer groups in a number of states are lobbying to legalize and regulate CPMs.[7] Our work highlights the value of information available in the market for healthcare services and could therefore inform new legislation. By measuring consumers' savvy when reading online reviews and establishing the predictive nature of doctor reviews, our work demonstrates that consumers are not blind to the quality of healthcare providers and indicates that perhaps legislation should lean more toward consumer choice than it has the past.

Limitations & Future Work

We plan to extend this work into other contexts and to use more robust methods to account for some of the limitations of this research. First, as noted in Essay 1, our experimental measures are self-reported. We would like to extend this work to measures of incentivized consumer choice. Another limitation is our restriction to services. While most of the theoretical literature focuses on credence services (also referred to as expert services), markets for credence goods also demonstrate information failures. For example, it is difficult for consumers to verify whether "organic" foods are actually organically produced, and consumers are often confused about the definition of the organic label they pay for (Harper and Makatouni 2002). It is currently unclear how online reviews or other consumer-generated

[7] c.f. Maryland: http://www.marylandfamiliesforsafebirth.org/legislation.html; Illinois: http://www.illinoismidwifery.org/blog/home; North Carolina Home Birth Freedom Act http://www.ncleg.net/gascripts/BillLookUp/BillLookUp.pl?Session=2013&BillID=H154

information could be used in markets like these; however, the changing nature of market regulation provides a strong motivation for future research. In at least one case, a state government has banned the undercover filming of food production on farms (Chappell 2014): rather than intervening to *increase* information in the market, government is intervening to *decrease* available information. We would like to study consumer-generated information in this type of context.

In Essay 2, we demonstrate the mathematical utility of online reviews in predicting medical board sanctions, and we discuss implications on consumer choice of doctor services. We note, however, that for a predictive model to have real impact on consumer choice, consumers must be able to understand and interpret the prediction results. As we found in Essay 1, and as is commonly known in the field of behavioral economics (c.f. Tversky and Kahneman 1986), consumers are not always adept at behaving as rational theories expect, and therefore they may be unable to use predictive information in an optimal way.

In the future, we hope to expand our work to develop a deeper and broader understanding of how consumers find and use information to make high-uncertainty decisions, how sellers respond to these information sources, and how platform-makers can elicit and present the most credible and valuable product-quality information. In the immediate future, we plan to study credence sellers' responses to reputational concerns that have arisen in response to online review platforms (e.g., post-sanction review impacts), which is a contentious issue in economic theory (Ely and Valimaki 2003; Grosskopf and Sarin 2010). We would also like to study how a

seller's signaling of reputational concerns (e.g., by participating on review platforms and frequently replying to reviews) may alter consumers' perception of the seller's credibility and the credibility of the seller's reviews. More generally, we are interested in continuing to develop understanding of how human bias affects the use of online information and an individual's choices and judgments.

Appendices

Table 14: "Other" category codes, examples of each, and percentage of data from Essay 1, Study 1

Other Category	**Example Snippet for Hair Stylists**	**Number of Snippets Coded (Percent Data)**
Reviewer's claim of expertise with service provider	"Adam has been cutting my hair for three years."	45 (2.6%)
Reviewer's claim of expertise with service category	"I've been to almost every hair salon in this city!"	49 (2.9%)
Coupon	"I used the Groupon."	10 (.59%)
Fixed Problem	"She fixed the messy highlights."	96 (5.6%)
Problem Description	"I went in for a cut, color, and blow dry."	68 (4.0%)
Recommendation	"You should try them!"	117 (6.9%)
Does not fit	"I have never written a review before, but We could not let my experience at this 'salon' slide."	153 (6.9%)
Total snippets coded as "Other"		538 (31.6%)

Table 15: Sample of state medical board mission statements

State	**Medical Board Mission Statement**
Maryland	The mission of the Board of Physicians is to assure quality health care in Maryland through the efficient licensure and effective discipline of health providers under its jurisdiction, by protecting and educating clients/customers and stakeholder, and enforcing the Maryland Medical Practice Act.
California	The mission of the Medical Board of California is to protect health care consumers through the proper licensing and regulation of physicians and surgeons and certain allied health care professions and through the vigorous, objective enforcement of the Medical Practice Act, and to promote access to quality medical care through the Board's licensing and regulatory functions.
Minnesota	The mission of the Minnesota Board of Medical Practice is to protect the public's health and safety by assuring that the people who practice medicine or as an allied health professional are competent, ethical practitioners with the necessary knowledge and skills appropriate to their title and role.
Texas	Safeguarding the public through professional accountability
Oklahoma	To promote the Health, Safety and Well-being of the citizens (patients) of Oklahoma by requiring a high level of qualifications, standards and continuing education for licenses regulated by Oklahoma Medical Board. To protect the on-going Health Safety and Well-being of the citizens (patients) of Oklahoma by investigating complaints, conducting public hearings, effectuating and monitoring disciplinary actions against any of the licensed professionals, while providing the licensee with proper due process and all rights afforded under the law. To provide any member of society upon request, a copy of the specific public records and information on any of the licensed professionals.

Table 16: Bases for medical board action and observable codes

Basis for Medical Board Action	Observable Code
Abandoned Medical Practice without Adequate Notice/Referrals	1
Abandoned Patient	1
Abuse of office/hospital staff	0
Abusive Billing Practices	0
Action by Hospital/Clinic/Professional Organization	0
Aiding or Abetting Practice without a License	0
Alcohol Abuse	0
Alcoholism	0
Allowing Physician Assistant to Perform Duties/Procedures not Appropriate to Practice	0
Allowing Staff to Perform Duties/Procedures without Appropriate Qualifications/Credentials	0
Allowing Unlicensed Person to Practice	0
Alteration/Falsification of Medical Record(s)	0
Alteration/Falsification of Research Findings	0
Assault	0
Attempting to Obtain a License by Misrepresentation	0
Battery	0
Breach of Confidentiality	0
Cheating	0
Chemical Abuse	0
Chemical Dependency	0
Competency Issue	0
Conduct Likely to Deceive or Defraud or Harm the Public	0
Conduct/Practice Which Is Or Might Be Harmful/Dangerous to the Health of the Patient/Public	0
Continuing Medical Education Violations	0
Controlled Substance Abuse	0
Controlled Substance Violations	0
Convicted of a Crime	0
Convicted of a Felony	0
Convicted of a Misdemeanor	0
Convicted of Assault	0
Convicted of Battery	0
Convicted of Crime of Moral Turpitude	0
Convicted of Criminal Sexual Conduct	0
Convicted of DUI/DWI	0
Convicted of Failing to Comply with Child Support Obligations	0
Convicted of Grand Larceny	0
Convicted of Homicide	0

Convicted of Larceny	0
Convicted of Manslaughter	0
Convicted of Money Laundering	0
Convicted of Murder	0
Convicted of Negligent Homicide	0
Convicted of Performing an Illegal Abortion(s)	0
Convicted of Procuring an Illegal Abortion(s)	0
Convicted of Public Lewdness	0
Convicted of Rape	0
Convicted of Receiving/Concealing Stolen Property	0
Convicted of Vehicular Homicide	0
Conviction for Patient Abuse or Neglect	0
Conviction Relating to Controlled Substances	0
Conviction Relating to Fraud	0
Conviction Relating to Health Care Fraud	0
Conviction Relating to Obstruction of an Investigation	0
Copying	0
Court Martial	0
Criminal Sexual Conduct	0
Default on Health Education Loan or Scholarship Obligations	0
Delinquent Taxes	0
Determination of Irregular Behavior	0
Dispensing Unlawfully	0
Due to Action Taken by Another Board/Agency	0
DUI/DWI	0
Examination Irregularities	0
Excessive Claims or Furnishing of Unnecessary or Substandard Items or Services	0
Excessive Prescribing	0
Excessive Treatment Not Warranted by Patient's Condition	0
Excessive/Inappropriate Use of Alcohol	0
Failure to Adequately Supervise	0
Failure to Adequately Supervise Medical Office Staff	0
Failure to Adequately Supervise Physician Assistant	0
Failure to Appear Before the Board as Directed	0
Failure to Appropriately Dispose of Controlled Substances/Drugs in Accordance with the Law	0
Failure to Appropriately Store Controlled Substances	0
Failure to Comply with Board Ordered Physical/Mental Evaluation	0
Failure to Comply with Child Support Obligations Established by Law	0
Failure to Comply with CME Requirements	0
Failure to Comply with Insurance Responsibility	0
Failure to Conform to Minimal Standards of Acceptable Medical Practice	0

Failure to Disclose Required Information	0
Failure to Examine Patient Prior to Initiation of Treatment	1
Failure to Examine/Evaluate Patient(s) in a Thorough Manner	0
Failure to Grant Immediate Access	0
Failure to Maintain Adequate Medical Records	0
Failure to Maintain Records of Prescribed/Dispensed Substances	0
Failure to Maintain/Submit CME Documentation	0
Failure to Meet Clerkship Requirements	0
Failure to Meet Education Requirements	0
Failure to Meet Examination Requirements	0
Failure to Meet Postgraduate Training Requirements	0
Failure to Meet Requirements	0
Failure to Notify Board of Address Change	0
Failure to Obtain Appropriate Consent	0
Failure to Pay CAT Fund Emergency Surcharge	0
Failure to Pay Income Tax	0
Failure to Provide Adequate Medical Coverage of Practice During Absence	1
Failure to Provide Appropriate Referrals	0
Failure to Provide Emergency/Timely Treatment	1
Failure to Provide Payment Information	0
Failure to Provide/Transfer Medical Records in a Timely Manner	1
Failure to Recognize Drug Seeking Behavior	0
Failure to Renew License	0
Failure to Renew License/Registration	0
Failure to Renew State Controlled Substance License	0
Failure to Report Adverse Actions Against Self in Accordance with Laws/Rules of the Board	0
Failure to Report Suspected Child Abuse	0
Failure to Respond to Request of the Board	0
Failure to Satisfy Terms of Prior Board Order	0
Failure to Supply Requested Information on Subcontractors and Suppliers	0
Failure to Take Corrective Action	0
False Advertising	0
Falsification of Licensure Application	0
Falsification of Scores	0
Falsification/Misrepresentation of Application Information	0
Felony Conviction Relating to Controlled Substance Violations	0
Felony Conviction Relating to Health Care Fraud	0
Fraud	0
Fraud	0
Fraud, Kickbacks and Other Prohibited Activities.	0
Fraudulent Billing Practices	0
Fraudulent Testimony as Medical Expert	0

Gross Negligence	0
Health-Related Problems	0
Homicide	0
Immediate Danger to the Public Health, Safety, or Welfare	0
Impairment	0
Imposition of a Civil Money Penalty or Assessment	0
Improper Management of Medical Records	0
Inappropriate Acquisition of Controlled Substances	0
Inappropriate Advertising	0
Inappropriate Prescribing	0
Inappropriate Treatment/Diagnosis	0
Income Tax Evasion	0
Income Tax Fraud	0
Incompetency	0
Individuals Controlling Sanctioned Entities.	0
Insurance Fraud	0
Intemperate Use of Alcohol	0
Internet Prescribing	0
Larceny	0
License Revocation or Suspension	0
Mail Fraud	0
Making A False/Fraudulent/Misleading Statement	0
Making or Assisting in Making Inappropriate Health Care Benefit Claims	0
Malpractice	1
Manslaughter	0
Medicaid Fraud	0
Medi-Cal Fraud	0
Medicare Fraud	0
Medicare/Medicaid Fraud	0
Mental Abuse of Patient	1
Mental Impairment	0
Mental-Health Related Problems	0
Misrepresentation of Medical Credentials/Qualifications	0
Misrepresentation of Possible Outcome/Complications of Treatment/Procedure	0
Misrepresentation on Document/Records	0
Moral Turpitude	0
Moral Unfitness	0
Morally Unfit to Practice Medicine	0
Negligence	0
Negligent Homicide	0
Not Applicable	0
Not Reported	0

Obtaining License by Fraudulent Misrepresentation	0
Overcharging for Copying/Providing Medical Records	0
Overutilization of Health Care Services	0
Patient Abuse	1
Performed Improper or Unnecessary Surgery	0
Performing an Illegal Abortion(s)	0
Physical Abuse of Patient	1
Physical Impairment	0
Physician-Patient Boundary Issues	1
Practicing Medicine While Under the Influence	1
Practicing Outside Scope of Medical Practice	0
Practicing the Profession Fraudulently	0
Practicing with Lapsed License	0
Practicing Without a License	0
Practicing Without Adequate Supervision	0
Prescribing and/or Dispensing Violation	0
Prescribing Drugs for Sexual Favors	1
Prescribing for Non-Therapeutic Purposes	1
Prescribing Unlawfully	0
Prescribing without Examination/Evaluation	1
Prescribing without Medical Indication/Need	0
Prescribing/Dispensing/Selling to Addicts	0
Pre-Signing of Prescription Blanks	0
PRO Recommendation	0
Procedural Violation	0
Procuring an Illegal Abortion(s)	0
Professional Misconduct	0
Program-Related Conviction	0
Public Lewdness	0
Receiving/Concealing Stolen Property	0
Security Violation	0
Sexual Boundary Issues	0
Sexual Misconduct	0
Substance Abuse	0
Suspension or Exclusion Under a Federal or State Health Care Program	0
Tax Fraud	0
Time Lapse Since Active Practice	0
Unable to Practice with Reasonable Skill and Safety	0
Unbecoming Conduct	0
Undetermined	0
Unethical Conduct	0
Unlawful Possession of Controlled Substances	0
Unprofessional Conduct	0

Verbal Abuse of Patient	1
Violation of a Prior Order of the Board	0
Violation of Consent Order	0
Violation of Interim Order	0
Violation of Prior Agreement	0
Violation of Probation	0
Violation of Statute or Rule of the Board	0
Violation of Stipulation and Order	0
Violation of Terms of Rehabilitation Stipulation and Order	0
Violation of Voluntary Affidavit Agreement	0
Willfully harassing, abusing, or intimidating a patient either physically or verbally	1
Willfully Making or Filing a False Report	0
Writing False or Fictitious Prescriptions	0
Wrong Site Procedure	1

References

Agarwal, R., and Dhar, V. 2014. "Big Data , Data Science , and Analytics : The Opportunity and Challenge for IS Research," *Information Systems Research* (25:3), pp. 443–448.

Allard, M., Wang, C., Kastanis, G., Pirone, C., Muruvanda, T., Strain, E., Timme, R., Payne, J., Luo, Y., Gonzalez-Escalona, N., ToroIbaceta, M., Ottesen, A., Melka, D., Evans, P., Musser, S., and Brown, E. 2015. "GenomeTrakr: A Pathogen Database to Build a Global Genomic Network for Pathogen Traceback and Outbreak Detection," in *International Association for Food Protection Annual Meeting*.

Andrews, M. 2008. "Rating Doctors : A Rank Practice ?," *U.S. News and Wolrd Report*, February 1.

Balafoutas, L., Beck, A., Kerschbamer, R., and Sutter, M. 2013. "What drives taxi drivers? A field experiment on fraud in a market for credence goods," *Review of Economic Studies* (80:3), pp. 876–891.

Bardhan, I., Oh, C., Zheng, E., and Kirksey, K. 2015. "Predictive Analytics for Readmission of Patients with Congestive Heart Failure," *Information Systems Research* (26:1), pp. 19–39.

Bateman, P., Gray, P., and Butler, B. S. 2011. "Research Note--The Impact of Community Commitment on Participation in Online Communities," *Information Systems Research* (22:4), pp. 841–854.

Beales, H., Craswell, R., and Salop, S. 1981. "The Efficient Regulation of Consumer Information," *Journal of Law and Economics* (XXIV:December).

Beck, A., Sutter, M., and Kerschbamer, R. 2010. "Guilt from Promise-Breaking and Trust in Markets for Expert Services: Theory and Experiment,"Institute for the Study of Labor.

Bloom, P. N., and Pailin, J. E. 1995. "Using information situations to guide marketing strategy," *Journal of Consumer Marketing* (12:2), pp. 19–27 (doi: 10.1108/07363769510084876).

Boudreau, M.-C., Gefen, D., and Straub, D. 2001. "Validation in Information Systems Research: A State-of-the-Art Assessment," *MIS Quarterly* (25:1), pp. 1–16.

Boyd-Graber, J., Mimno, D., and Newman, D. 2014. "Care and Feeding of Topic

Models: Problems, Diagnostics, and Improvementes," in *Handbook of Mixed Membership Models and Their Applications*A. E.M., D. M. Blei, E. A. Erosheva, and S. E. Fienberg (eds.), Boca Raton: CRC Press.

Cain, D. M., Loewenstein, G., and Moore, D. a. 2005. "The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest," *The Journal of Legal Studies* (34:1), pp. 1–25 (doi: 10.1086/426699).

Chappell, B. 2014. "Bill That Bans Undercover Filming At Farms Enacted In Idaho," *NPR*.

Cheung, M. Y., Luo, C., Sia, C. L., and Chen, H. 2009. "Credibility of Electronic Word-of-Mouth: Informational and Normative Determinants of On-line Consumer Recommendations," *International Journal of Electronic Commerce* (13:4), pp. 9–38 (doi: 10.2753/JEC1086-4415130402).

Chevalier, J., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research* (41:3), pp. C01–354 (doi: 10.1509/jmkr.43.3.345).

Clemons, E., Gao, G. G., and Hitt, L. 2006. "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry," *Journal of Management Information Systems* (23:2), pp. 149–171 (doi: 10.2753/MIS0742-1222230207).

Dafny, L., and Dranove, D. 2008. "Do report cards tell consumers anything they don't know already? The case of Medicare HMOs.," *The Rand journal of economics* (39:3), pp. 790–821.

Daniel, S., Agarwal, R., and Stewart, K. 2006. "An Absorptive Capacity Perspective of Open Source Software Development Group Performance," in *International Conference on Information Systems*.

Darby, M. R., and Karni, E. 1973. "Free competition and the optimal amount of fraud," *Journal of law and economics* (16:1), JSTOR, pp. 67–88.

Decker, R., and Trusov, M. 2010. "Estimating aggregate consumer preferences from online product reviews," *International Journal of Research in Marketing* (September), pp. 1–32.

Ding, Y., and Simonoff, J. S. 2010. "An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data," *Journal of Machine Learning Research* (11), pp. 131–170.

Dranove, D., and Jin, G. Z. 2010. "Quality Disclosure and Certification : Theory and

Practice," *Journal of Economic Literature* (48:4), pp. 935–963.

Dranove, D., Kessler, D., McClellan, M., and Satterthwaite, M. 2003. "Is More Information Better? The Effects of 'Report Cards' on Health Care Providers," *Journal of Political Economy* (111:3), pp. 555–588 (doi: 10.1086/374180).

Dranove, D., and Sfekas, A. 2008. "Start spreading the news: a structural estimate of the effects of New York hospital report cards.," *Journal of Health Economics* (27:5), pp. 1201–7 (doi: 10.1016/j.jhealeco.2008.03.001).

Dulleck, U., and Kerschbamer, R. 2006. "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods," *Journal of Economic Literature* (44:1), pp. 5–42 (doi: 10.1257/002205106776162717).

Dulleck, U., Kerschbamer, R., and Sutter, M. 2011. "The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition," *The American Economic Review* (101:2), pp. 526–555 (doi: 10.1257/aer.101.2.526).

ElBoghdady, D. 2012. "Some doctors try to squelch online reviews," *The Washington Post*, Washington, DC, January 28 (available at http://www.washingtonpost.com/business/economy/some-doctors-try-to-squelch-online-reviews/2011/11/29/gIQA2KQhYQ_story.html; retrieved February 16, 2012).

Ely, J., and Valimaki, J. 2003. "Bad Reputation," *The Quarterly Journal of Economics* (118:3), pp. 785–814 (doi: 10.1162/00335530360698423).

Emons, W. 1997. "Credence Goods and Fraudulent Experts," *The RAND Journal of Economics* (28:1), p. 107 (doi: 10.2307/2555942).

Fawcett, T. 2006. "An introduction to ROC analysis," *Pattern Recognition Letters* (27), pp. 861–874.

Fenton, J. J., Jerant, A., Bertakis, K., and Franks, P. 2012. "The Cost of Satisfaction," *Archives of Internal Medicine* (172:5), pp. 405–411 (doi: 10.1001/archinternmed.2011.1662).

Food and Drug Administration. 2013. "Structured Approach to Benefit-Risk Assessment in Drug Regulatory Decision-Making.,"

Ford, G., Smith, D. B., and Swasy, J. L. 1990. "Consumer skepticism of advertising claims: Testing hypotheses from economics of information," *Journal of Consumer Research* (16:4), JSTOR, pp. 433–441.

Ford, G., Smith, D., and Swasy, J. L. 1988. “An Empirical Test of the Search , Experience and Credence Attributes Framework,” *Advances in Consumer Research* (15).

Fox, S., and Jones, S. 2009. “The social life of health information,” *Pew Internet & Americal Life Project*.

Galetzka, M., Verhoeven, J. W. M., and Pruyn, A. T. H. 2006. “Service validity and service reliability of search, experience and credence services: A scenario study,” *International Journal of Service Industry Management* (17:3), pp. 271–283 (doi: 10.1108/09564230610667113).

Gao, G. G., Greenwood, B. N., Agarwal, R., and Mccullough, J. 2011. “A Digital Soapbox ? The Information Value of Online Physician Ratings.,”

Gao, G. G., McCullough, J., Agarwal, R., and Jha, A. 2012. “Online physician ratings by patients: A changing landscape,” *Jounral of Medical Internet Research* (14:1), p. e38.

Grosskopf, B., and Sarin, R. 2010. “Is Reputation Good or Bad? An Experiment,” *The American Economic Review* (100:December), pp. 2187–2204.

Grossman, S. 1981. “The informational role of warranties and private disclosure about product quality,” *Journal of law and economics* (24:3), pp. 461–483.

Gruber, J., Kim, J., and Mayzlin, D. 1999. “Physician fees and procedure intensity: the case of cesarean delivery.,” *Journal of health economics* (18:4), pp. 473–90.

Hahn, R., and Hird, J. 1991. “Costs and Benefits of Regulation: Review and Synthesis, The,” *Yale J. on Reg.* (1).

Hamilton, R. W., Schlosser, A., and Chen, Y.-J. 2015. “Who’s Driving This Conversation? Systematic Biases in the Content of Online Consumer Discussions,”Georgetown University.

Hamilton, R. W., and Thompson, D. V. 2007. “Is There a Substitute for Direct Experience? Comparing Consumers’ Preferences after Direct and Indirect Product Experiences,” *Journal of Consumer Research* (34:December), pp. 546–555.

Han, S., Mankad, S., Gavirneni, N., and Verman, R. 2016. “What Guests Really Think of Your Hotel : Text Analytics of Online Customer Reviews,” *Cornell Hostpirality Report* (16:2), pp. 3–17.

Hanley, J., and McNeil, B. 1982. “The Meaning and Use of the Area under a

Receiving Operating Characteristic (ROC) Curve," *Radiology* (143:1), pp. 29–36.

Hao, Y., Ye, Q., Li, Y., and Cheng, Z. 2010. "How does the Valence of Online Consumer Reviews Matter in Consumer Decision Making? Differences between Search Goods and Experience Goods," in *Proceedings of the 43rd Hawaii International Conference on System Sciences - 2010*, pp. 1–10.

Harper, G. C., and Makatouni, A. 2002. "Consumer perception of organic food production and farm animal welfare," *British Food Journal* (104:3/4/5), MCB UP Ltd, pp. 287–299 (doi: 10.1108/00070700210425723).

Harris, K., and Buntin, M. 2008. "Choosing a health care provider: the role of quality information," *Policy*.

Hastings, J., and Weinstein, J. 2008. "Information, school choice, and academic achievement: Evidence from two experiments," *The Quarterly Journal of Economics* (123:4), pp. 1373–1414.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., and Gremler, D. D. 2004. "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?," *Journal of Interactive Marketing* (18:1), pp. 38–52 (doi: 10.1002/dir.10073).

Huang, P., Lurie, N. H., and Mitra, S. 2009. "Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods," *Journal of Marketing* (73:2), pp. 55–69 (doi: 10.1509/jmkg.73.2.55).

Jain, S. 2010. "Googling ourselves--what physicians can learn from online rating sites.," *The New England Journal of Medicine* (362:1), pp. 6–7 (doi: 10.1056/NEJMp0903473).

Jensen, M. L., Averbeck, J. M., Zhang, Z., and Wright, K. B. 2013. "Credibility of Anonymous Online Product Reviews: A Language Expectancy Perspective," *Journal of Management Information Systems* (30:1), pp. 293–324 (doi: 10.2753/MIS0742-1222300109).

Jiménez, F. R., and Mendoza, N. A. 2013. "Too Popular to Ignore: The Influence of Online Reviews on Purchase Intentions of Search and Experience Products," *Journal of Interactive Marketing* (27:3), pp. 226–235.

Jin, G. Z., and Leslie, P. 2003. "The Effect Of Information On Product Quality: Evidence From Restaurant Hygiene Grade Cards," *Quarterly Journal of Economics* (May), pp. 409–451.

Joskow, P., and Noll, R. 1981. “Regulation in theory and practice: an overview,” *Studies in public regulation* (I), pp. 1–78.

Joskow, P., and Rose, N. 1989. “The effects of economic regulation,” in *Handbook of Industrial Organization, Volume II*R. Schmalenseee and R. D. Willig (eds.), Elsevier Science Publishers.

Jovanovic, B. 1982. “Truthful Disclosure of Information,” *Bell Journal of Economics* (13:1), The RAND Corporation, pp. 36–44.

Kang, J. S., Kuznetsova, P., Choi, Y., and Luca, M. 2013. “Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. R. I. 2012. “Leakage in Data Mining : Formulation , Detection , and Avoidance,” (6:4), pp. 1–21 (doi: 10.1145/2382577.2382579).

Kempf, D. S., and Smith, R. E. 1998. “Consumer Processing of Product Trial and the Influence of Prior Advertising: A Structural Modeling Approach,” *Journal of Marketing Research* (35:3), p. 325 (doi: 10.2307/3152031).

Kerschbamer, R., Dulleck, U., and Sutter, M. 2009. “The Impact of Distributional Preferences on The Impact of Distributional Preferences on (Experimental) Markets for Expert Services.,”

Kim, D., and Benbasat, I. 2006. “The Effects of Trust-Assuring Arguments on Consumer Trust in Internet Stores: Application of Toulmin’s Model of Argumentation,” INFORMS.

Kraut, R. E., and Resnick, P. 2010. “Encouraging contribution to online communities,” in *Evidence-based social design: Mining the social sciences to build successful online communities.*, Cambridge, MA: MIT Press.

Langer, E. J., Blank, A., and Chanowitz, B. 1978. “The Mindlessness of Ostensibly Thoughtful Action: The Role of ‘Placebic’ Information in Interpersonal Interaction,” *Journal of Personality and Social Psychology* (36:6), pp. 635–642 (doi: 10.1037/0022-3514.36.6.635).

Larrick, R. P., and Soll, J. B. 2006. “Intuitions About Combining Opinions: Misappreciation of the Averaging Principle,” *Management Science* (52:1), pp. 111–127 (doi: 10.1287/mnsc.1050.0459).

Lim, B. C., and Chung, C. M. Y. 2011. “The Impact of Word-of-Mouth

Communication on Attribute Evaluation," *Journal of Business Research* (64:1), pp. 18–23 (doi: 10.1016/j.jbusres.2009.09.014).

Liu, T. 2011. "Credence Goods Markets With Conscientious and Selfish Experts," *International Economic Review* (52:1), Wiley Online Library, pp. 227–244.

Lu, S. F. 2012. "Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes," *Journal of Economics & Management Strategy* (31:3), pp. 673–705.

Lu, S. F., and Rui, H. 2015. "Can We Trust Online Physician Ratings ? Evidence from Cardiac Surgeons in Florida," in *48th Hawaii International Conference on System Sciences*, pp. 2876–2885 (doi: 10.1109/HICSS.2015.348).

Luca, M. 2011. "Reviews, Reputation, and Revenue: The Case of Yelp.Com," *Harvard Business School - Negotiations, Organizations & Markets Unit*.

Luxford, K. 2012. "What does the patient know about quality ?," *International Journal for Quality in Health Care* (24:5), pp. 439–440.

Manning, C. D., Raghavan, P., and Schutze, H. 2009. *Scoring, term weighting, and the vector space model Introduction to Information Retrieval*, Boston, MA: Cambridge University Press.

Mayer, R. C., and Davis, J. H. 1999. "The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-experiment," *Journal of Applied Psychology* (84:1), pp. 123–136 (doi: 10.1037//0021-9010.84.1.123).

Mayfield, E., and Rosé, C. P. 2013. "LightSIDE: Open Source Machine Learning for Text," in *Handbook of automated essay evaluation: Current applications and new directions.*

Mimra, W., Rasch, A., and Waibel, C. 2012. "Reputation in credence goods markets: Experimental evidence,"Center of Economic Research at ETH Zurich (Vol. 49).

Moe, W. W., and Trusov, M. 2011. "The Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research* (48:3), pp. 444–456 (doi: 10.1509/jmkr.48.3.444).

Moorman, C., Ferraro, R., and Huber, J. 2012. "Unintended nutrition consequences: firm responses to the nutrition labeling and education act," *Marketing Science* (March 2014).

Mudambi, S. M., and Schuff, D. 2010. "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly* (34:1), pp. 185–

200.

Neal, D., and Schanzenbach, D. 2010. “Left behind by design: Proficiency counts and test-based accountability,” *Review of Economics and Statistics* (92:2), pp. 263–283 (doi: 10.1162/rest.2010.12318).

Nelson, P. 1970. “Information and Consumer Behavior,” *The Journal of Political Economy* (78:2), JSTOR, pp. 311–329.

Ohanian, R. 1990. “Construction and validation of a scale to measure celebrity endorsers’ perceived expertise, trustworthiness, and attractiveness,” *Journal of Advertising* (19:3), pp. 39–52.

Park, D.-H., and Kim, S. 2008. “The Effects of Consumer Knowledge on Message Processing of Electronic Word-of-mouth via Online Consumer Reviews,” *Electronic Commerce Research and Applications* (H. Österle, J. Schelp, and R. Winter, eds.) (7:4), Elsevier B.V., pp. 399–410 (doi: 10.1016/j.elerap.2007.12.001).

Parker, D., and Kirkpatrick, C. 2012. “The Economic Impact of Regulatory Policy: A Literature REview of Quantitative Evidence,”Organisation for Economic Co-operation and Development Expert Paper No. 3, August 2012.

Parker, G. G., and Van Alstyne, M. W. 2005. “Two-Sided Network Effects: A Theory of Information Product Design,” *Management Science* (51:10), pp. 1494–1504 (doi: 10.1287/mnsc.1050.0400).

Pavlou, P. A., and Fygenson, M. 2006. “Understanding and Predicting Electronic Commerce Adoption: An Extension of the Theory of Planned BehaviorNo Title,” *MIS Quarterly* (30:1), pp. 115–143.

Pencina, M. J., D’Agostino, R. S., D’Agostino, R. J., and Vasan, R. 2008. “Evaluating the Added Predictive Ability of a New Marker: From Area Under the ROC Curve to Reclassification and Beyond,” (April 2007), pp. 157–172 (doi: 10.1002/sim).

Pornpitakpan, C. 2004. “The Persuasiveness of Source Credibility: A Critical Review of Five Decades’ Evidence,” *Journal of Applied Social Psychology* (34:2), Wiley Online Library, pp. 243–281.

Racherla, P., Mandviwalla, M., and Connolly, D. 2012. “Factors affecting consumers’ trust in online product reviews,” *Jounral of Consumer Behavior* (11), pp. 94–104 (doi: 10.1002/cb).

Rasch, A., and Waibel, C. 2012. “What Drives Fraud in a Credence Goods Market?—

Evidence from a Quasi Field Experiment,"CER-ETH Economics working paper (Vol. 49).

Robb, M. A., Racoosin, J. A., Sherman, R. E., Gross, T. P., Ball, R., Reichman, M. E., Midthun, K., and Woodcock, J. 2012. "The US Food and Drug Administration's Sentinel Initiative: Expanding the horizons of medical product safety," *Pharmacoepidemiology and Drug Safety* (21:September 2007), pp. 9–11 (doi: 10.1002/pds).

Rothschild, M., and Stiglitz, J. 1992. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," in *Foundations of Insurance Economics*G. Dionne and S. E. Harrington (eds.), Springer Netherlands, pp. 355–375.

Rust, R., and Cooil, B. 1994. "Reliability measures for qualitative data: Theory and implications," *Journal of Marketing Research* (31:February 1994), pp. 1–14.

Salop, S. 1977. "The noisy monopolist: imperfect information, price dispersion and price discrimination," *The Review of Economic Studies* (44:3), pp. 393–406.

Schlosser, A. E., White, T. B., and Lloyd, S. M. 2006. "Converting Web Site Visitors into Buyers: How Web Site Investment Increases Consumer Trusting Beliefs and Online Purchase Intentions," *Journal of Marketing* (70:2), pp. 133–148.

Schneider, H. 2009. "Agency Problems and Reputation in Expert Services: Evidence from Auto Repair," *Journal of Industrial Economics* (60:3), pp. 406–433 (doi: http://ssrn.com/abstract=1022204).

Scholle, S. H., Roski, J., Dunn, D. L., Adams, J. L., Dugan, D. P., Pawlson, L. G., and Kerr, E. a. 2009. "Availability of data for measuring physician quality performance.," *The American Journal of Managed Care* (15:1), pp. 67–72.

Schwartz, A., and Wilde, L. 1978. "Intervening in markets on the basis of imperfect information: A legal and economic analysis," *U. Pa. L. Rev.*

Shmueli, G. 2010. "To Explain or to Predict?," (25:3), pp. 289–310 (doi: 10.1214/10-STS330).

Shmueli, G., and Koppius, O. R. 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553–572.

Spence, M. 1973. "Job Market Signaling," *The Quarterly Journal of Economics* (87:3), Oxford University Press, p. 355 (doi: 10.2307/1882010).

Srinivasan, S. S., and Till, B. D. 2002. "Evaluation of Search, Experience and

Credence attributes: Role of Brand Name and Product Trial," *Journal of Product & Brand Management* (11:7), MCB UP Ltd, pp. 417–431 (doi: 10.1108/10610420210451616).

Stewart, K. J. 2003. "Trust Transfer on the World Wide Web," *Organization Science* (14:1), pp. 5–17 (doi: 10.1287/orsc.14.1.5.12810).

Strack, F. 1992. ""Order Effects'' in Survey Research: Activation and Information Functions of Preceding Questions," in *Context Effects in Social and Psychological Research* N. Schwarz and S. Sudman (eds.), New York, NY: Springer New York, pp. 23–34 (doi: 10.1007/978-1-4612-2848-6_3).

Straub, D., Boudreau, M.-C., and Gefen, D. 2004. "Validation Guidelines for IS Positivist Research," *Communications of the Association for Information Systems* (13:24), pp. 380–427.

Sun, M. 2012. "How does the variance of product ratings matter?," *Management Science* (March 2014).

Thaler, R. H., and Sunstein, C. R. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Penguin Books.

Toulmin, S. 1958. *The Uses of Argument*, Cambridge, Eng: University Press.

Tversky, A., and Kahneman, D. 1986. "Rational Choice and the Framing of Decisions," *Journal of business*, JSTOR, pp. 251–278.

Viscusi, W., Magat, W., and Huber, J. 1986. "Informational regulation of consumer health risks: an empirical evaluation of hazard warnings," *The Rand Journal of Economics* (17:3).

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., and Dredze, M. 2014. "A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews," *J Am Med Inform Assoc* (21), pp. 1098–1103 (doi: 10.1136/amiajnl-2014-002711).

White, W. 2004. "Market Forces, Competitive Strategies, and Health Care Regulation," *U. Ill. L. Rev.*, pp. 137–166.

Wolinsky, A. 1993. "Competition in a Market for Informed Experts' Services," *The RAND Journal of Economics* (24:3), p. 380 (doi: 10.2307/2555964).