Kerschbamer, Rudolf; Neururer, Daniel; Sutter, Matthias

Working Paper

Credence Goods Markets and the Informational Value of New Media: A Natural Field Experiment

IZA Discussion Papers, No. 12184

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Kerschbamer, Rudolf; Neururer, Daniel; Sutter, Matthias (2019) : Credence Goods Markets and the Informational Value of New Media: A Natural Field Experiment, IZA Discussion Papers, No. 12184, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/196682

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12184

Credence Goods Markets and the Informational Value of New Media: A Natural Field Experiment

Rudolf Kerschbamer
Daniel Neururer
Matthias Sutter

FEBRUARY 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12184

Credence Goods Markets and the Informational Value of New Media: A Natural Field Experiment

Rudolf Kerschbamer
University of Innsbruck

Daniel Neururer
University of Innsbruck

Matthias Sutter
MPI for Research on Collective Goods, University of Cologne, University of Innsbruck and IZA

FEBRUARY 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Credence Goods Markets and the Informational Value of New Media: A Natural Field Experiment*

Credence goods markets are characterized by pronounced informational asymmetries between consumers and expert sellers. As a consequence, consumers are often exploited and market efficiency is threatened. However, in the digital age, it has become easy and cheap for consumers to self-diagnose their needs using specialized webpages or to access other consumers' reviews on social media platforms in search for trustworthy sellers. We present a natural field experiment that examines the causal effect of information acquisition from new media on the level of sellers' price charges for computer repairs. We find that even a correct self-diagnosis of a consumer about the appropriate repair does not reduce prices, and that an incorrect diagnosis more than doubles them. Internet ratings of repair shops are a good predictor of prices. However, the predictive valued of reviews depends on whether they are judged as reliable or not. For reviews recommended by the platform Yelp we find that good ratings are associated with lower prices and bad ratings with higher prices, while non-recommended reviews have a clearly misleading effect, because non-recommended positive ratings increase the price.

JEL Classification: C93, D82

Keywords: credence goods, fraud, information acquisition, internet, field experiment

Corresponding author:
Matthias Sutter
Max Planck Institute for Research on Collective Goods
Kurt Schumacher Strasse 10
D-53113 Bonn
Germany
E-mail: matthias.sutter@coll.mpg.de

* We benefitted from comments by Loukas Balafoutas, Niall Flynn, Ben Greiner, Axel Ockenfels, Henry Schneider, Marco Schwarz, Christian Waibel and seminar participants at UC San Diego, UC Riverside and the universities of Amsterdam, Copenhagen, Dijon, Göteborg, Göttingen, Jena, Karlsruhe, Mannheim, and Tübingen. Brian Cooper helped editing the document. Financial support from the Austrian Science Fund (FWF) through special research area grant SFB F63, as well as through grant numbers P26901 and P27912, and from the German Science Foundation (DFG) through the Excellence Cluster ECONtribute: Markets & Public Policy is gratefully acknowledged.

1. Introduction

Markets for credence goods (Darby and Karni, 1973; Dulleck and Kerschbamer, 2006; Huck et al., 2016a) are ubiquitous in daily life. They include, among others, markets for health care, repair and legal services, as well as financial advice and fund management, with all of these markets having a huge size in the overall economy.[1] The key feature of these markets is the informational asymmetry between expert sellers and consumers: Doctors, mechanics, and legal or financial experts are typically much better informed about the quality of a good, service or asset that fits a consumer's needs best than patients, clients or private investors. Consumers are often even unable to judge *ex post* whether a particular provision was appropriate or not.[2]

The pronounced informational asymmetries present on markets for credence goods create strong material incentives for expert sellers to cheat on consumers, particularly through overprovision or overcharging (Dulleck and Kerschbamer, 2006).[3] Overprovision means that expert sellers provide a higher quality (or quantity) than the level that would have maximized the gains from trade. This creates an immediate inefficiency since the additional benefits to the consumer from the higher quality (or quantity) are lower than the additional costs. An example for overprovision is a car mechanic replacing a filter when cleaning it would have been sufficient. Overcharging refers to experts charging for more than they have actually provided – like a car mechanic putting a new filter on the bill when actually he only cleaned the old filter. Overcharging can also lead to inefficiencies in the long run if the fear of getting overcharged deters consumers from trading on credence goods markets in the future. Such a process could ultimately even lead to the breakdown of the market (similar to Akerlof's, 1970, analysis of lemons markets).

The fact that the superior information of sellers threatens the efficiency of credence goods markets and puts consumers at the risk of exploitation raises the question how to contain such negative effects of informational asymmetries. One straightforward approach would be to narrow down or even close the information gap between sellers and consumers. In fact, modern

[1] For instance, health care expenditures alone account for about 10% of GDP in the OECD-countries (www.oecd-library.org). The finance sector represents 9% of worldwide GDP (see *The Economist*, 2014: http://www.economist.com/news/finance-and-economics/21604574-new-paper-shows-industrys-take-has-been-rising-counting-cost-finance), and repair services generate more than 100 billion Euro per annum in Europe alone (ec.europa.eu/eurostat). Links accessed on 10 January 2019.

[2] Somewhat related to the work on credence goods markets are papers by Huck et al. (2012, 2016b) who study the provision of experience goods. These goods (like wine) have characteristics that are unobservable for the consumer *ex ante*, but the quality is revealed after buying or consuming them and therefore consumers can judge *ex post* whether they received the quality that yields the highest gains from trade or not. The latter is not possible with credence goods (unless underprovision occurs).

[3] Gneezy (2005) shows that cheating behavior (in a sender-receiver game) depends in a very systematic way on the material incentives for cheating. Laboratory evidence for such a relationship in a credence goods setting is presented in Beck et al. (2013).

communication technologies and social media have made it much easier and cheaper for consumers to inform themselves. Yet, it is by no means clear whether the information available on the internet actually helps consumers and, if so, by how much. For this reason, we present a novel natural field experiment (List and Rasul, 2011) in the computer repair market, which examines the causal effects of information retrieved from the internet on the extent of fraudulent behavior of sellers.[4]

Why would modern communication technologies and social media potentially be able to affect the level of honest provision, and thus efficiency, on credence goods markets? One can think of two main channels that can help consumers on credence goods markets to contain exploitation through sellers. The first channel works through specialized internet pages that allow individual consumers to self-diagnose their needs, thereby reducing the extent of the informational asymmetry between the seller and the consumer.[5] The second channel works through internet reviews by other consumers which can help to identify expert sellers who provide appropriate quality at reasonable prices, thus limiting overprovision and overcharging.

There are numerous examples for the first channel. For instance, there are several webpages that allow consumers to self-diagnose the problem in case a computer can no longer be booted. In such situations, the computer issues a characteristic sequence of beeps, based on which specialized webpages can facilitate the identification of the source of the problem.[6] This reduces the degree of asymmetric information on the appropriate repair and, therefore, might help computer owners. In markets for financial advice, regulators hope that automated robo-advisers provide cheap and unbiased advice to low income investors, thus reducing the informational advantage of human financial advisors over their consumers (D'Acunto et al., 2018). In markets for health care services, some webpages allow a patient to enter his or her symptoms and then generate a diagnosis. Also, patients can upload X-rays to internet portals to get an opinion about their health problems (as implemented in Gottschalk et al., 2018)[7], thereby helping them to avoid useless treatments that tackle the wrong symptoms. As a final example, smart-phone applications like Google Maps have made it very easy – and practically costless – for taxi

[4] Seen from a broader perspective, our paper relates to the newly emerging field of forensic economics (Zitzewitz, 2012) and to the literature on what drives individual propensities to act morally or to cheat on others (e.g., Cappelen et al., 2017; Gneezy et al., 2018; Abeler et al., 2019). Yet, these strands of the literature have not investigated credence goods markets (despite the immense scope of such markets in modern economies).

[5] Of course, also before the advent of the internet, reducing the degree of asymmetry in the information of sellers and consumers was possible for consumers by searching for offline information. The internet, however, has made it so much easier to acquire information cheaply and almost instantaneously so that the informational asymmetries might get reduced to an extent not possible before new media revolutionized the access to information.

[6] For Lenovo machines, for instance, this is the following (German) page: https://support.lenovo.com/at/en/solutions/ht035729 (accessed on 10 January 2019).

[7] See also, for instance, https://www.netdoktor.de/symptom-checker/ or https://www.secondopinions.com (accessed on 10 January 2019).

passengers to find the shortest route to a given destination in an unknown city. This might help them to avoid being taken on unreasonable detours – a classic form of overprovision in such markets (Balafoutas et al., 2013, 2017).

The second channel through which modern technologies might help consumers on credence goods markets is the plethora of rating platforms (like Yelp or the one on Google) on which consumers give feedback and rate sellers of different types of goods and services. Some of these platforms refer to particular types of credence goods providers, such as physicians, repair shops, or lawyers.[8] With regards to taxi drivers, ratings of these credence goods providers are already inbuilt in Uber's services, for instance, as a quality control measure. If reliable, the information contained on rating platforms might also directly help consumers by guiding them to trustworthy expert sellers. Through this channel, rating platforms could increase the trade volume and thus efficiency on credence goods markets.[9]

So far, there is no controlled evidence in support of the conjecture that modern technologies actually help consumers to receive appropriate provision of goods and services and to get overcharged less than when these technologies are not used. Referring to the first channel of information for consumers discussed above, it seems plausible that consumers receive more appropriate service or better prices if they demonstrate to expert sellers that they have acquired a good knowledge of their needs. Several studies have shown that a lack of knowledge is clearly disadvantageous for consumers, but it is as yet unclear whether consumers benefit from receiving and revealing additional – but in most cases also noisy – information about their needs, compared to a situation where consumers do not reveal anything. Balafoutas et al. (2013, 2017) show that taxi passengers who reveal that they are unfamiliar with the optimal route to their destination or with the tariff system in the respective city are more likely to be taken on detours and to be overcharged than passengers who simply state the requested destination. Gottschalk et al. (2018) report that patients who are perceived as less informed about the proper treatment at a dentist are more likely to be overtreated (with respect to the recommended treatment). All of these papers compare binary information levels and investigate whether changing the information level of consumers has an effect on the behavior of sellers. We

[8] See, for example, https://www.jameda.de or https://lawyers.com or, more generally, https://www.yelp.com (accessed on 10 January 2019).

[9] There is a large literature on how ratings of sellers on internet trading platforms affect the behavior of consumers (see, e.g., Bolton et al., 2004, Bohnet and Huck, 2004, Grosskopf and Sarin, 2010, Bolton et al., 2013; Huck et al., 2016b; Bolton et al., 2019). This literature typically investigates sales offers on trading platforms like eBay or Amazon on which the price for the good is *not* unknown to consumers *before* an interaction. Rather, in these cases the quality of the product is unknown ex ante. In our field experiment, the price of an interaction is part of the moral hazard problem. More importantly, we analyze a credence goods market where consumers are not even able to judge *ex post* whether they have received the appropriate quality, which distinguishes our work from the aforementioned papers.

examine a broader question, namely whether it is at all useful for consumers to acquire additional information on the internet that – in the context of credence goods markets – will almost always remain noisy and will not transform a consumer into an expert about the credence good. In the case of computer repairs in our field experiment, a webpage will facilitate a self-diagnosis, but the consumer will hardly be able to judge whether the self-diagnosis is correct or not. So, the general question is whether any self-diagnosis through sources from the internet – be it correct or not – will benefit the consumer. To address this question it is necessary to compare three conditions – one in which the consumer has not acquired a diagnosis through sources from the internet, one in which the consumer has acquired an incorrect signal and one in which the consumer has acquired a correct signal. To the best of our knowledge no previous study has made such a comparison.

The second channel – internet reviews – which consumers might resort to for information about expert sellers need not be helpful for consumers either. First, and foremost, it is less than obvious that reviews posted on rating platforms contain useful information as soon as credence goods are concerned. This is due to the fact that in credence goods markets consumers are typically not even able to judge *ex post* whether they were provided the good or service that maximized the gains from trade. For this reason, the ratings that users may post on rating platforms are not necessarily a reliable source of information for actual seller behavior. Second, sellers might be able to manipulate or fake the ratings themselves or commission benevolent ones, thus reducing the information content of internet ratings further (see Ockenfels and Resnick, 2012).[10]

To examine the influence of both channels through which consumers in credence goods markets can acquire information, we ran a natural field experiment in computer repair shops in Germany. In a first wave of the experiment, we manipulated completely refurbished computers in a controlled way and then handed them in for repair. We used three treatments: (1) a control treatment where we do not mention any possible source of the problem; (2) a treatment where the owner of the computer mentions a correct self-diagnosis of the problem; and (3) a treatment where the diagnosis offered by the consumer is wrong. We find that mentioning a correct self-diagnosis about the appropriate repair does not reduce the repair price in comparison to a situation where the computer owner simply asks for a repair, and mentioning an incorrect diagnosis more than doubles the average price. This constitutes our first main result. Since the signal generated by specialized diagnosis software in the internet is almost always noisy and

[10] As early as 2012, the *New York Times* wrote an article about commissioned reviews of all sorts of products to attract the attention of consumers. See https://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html (accessed on 10 January 2019).

since consumers cannot distinguish between a correct and an incorrect diagnosis, an implication of our first main result is that acquiring and revealing a noisy self-diagnosis is a costly mistake for consumers in markets for credence goods.

Only after running the first wave of our field experiment did we collect internet ratings of the shops that we had visited, finding that negative ratings are significantly associated with higher prices, while positive ratings have only a mild effect on reducing prices. Based on this *ex post* result, we ran a second wave of the field experiment in a new city. To assess the empirical relation between ratings and repair prices, we selected the shops with the highest number of internet reviews for inclusion, and we expected *ex ante* significantly lower prices in the better rated shops than in the shops with worse ratings. This is exactly our second main result, with prices about 50% higher in the shops with worse ratings.

Finally, we dug deeper into the predictive power of internet ratings by relying on a classification of ratings into recommended ones and non-recommended ones, where the latter are considered by the respective platform as less reliable. We find that for recommended ratings the relationship between ratings and repair prices is as expected: Shops with more positive ratings have significantly lower repair prices and shops with more negative ratings have significantly higher prices. Yet, when we look at non-recommended ratings, we see that more positive ratings predict *higher* (rather than lower) prices. Together these findings suggest that internet ratings may be a cursed blessing, as it is crucial whether ratings are reliable or somehow manipulated (by expert sellers or commissioned reviewers). This constitutes our third main result.

In the following, we present our experimental design of the first wave in section 2, and the results of it in section 3. The second wave's design and results are presented in sections 4 and 5. Section 6 concludes the paper.

2. Self-diagnosis through webpages – Wave 1 of the field experiment

We conducted the first wave of our field experiment in several German cities – Bonn, Cologne, Düsseldorf, Leverkusen, and Munich. In order to run it, we bought 12 identical, completely refurbished and perfectly working laptops (see Appendix B for the detailed specification). In each of the computers, we removed the random-access memory (RAM) modules slightly from their slots. Loose RAM modules are not an exotic problem; rather, it can occur easily if a laptop drops on the floor, for instance. A consequence of the loose RAM modules is that the computer is prevented from booting, causing a black screen and a distinct

acoustic error message. Lenovo has a webpage that allows the inferring of the most likely problem conditional on the acoustic messages.[11] In our case, the page suggests, as the most likely cause, a problem with the RAM modules, and as an alternative, but less likely, problem an issue with the main board. According to our university's IT department, a competent repair shop should be able to diagnose and solve this problem correctly within 10-15 minutes. For a consumer, it remains ambiguous, however, what the real problem is, for which reason appropriate service and pricing are hard to identify.

The fact that the problem can easily be diagnosed and repaired by a computer shop is an important feature of our experiment because our primary research interest is in intentional fraud and not in incompetence. Moreover, the estimated repair time of 10 or 15 minutes represents another important feature of our experiment, because most shops in Germany charge their working time in intervals of 30 minutes. This means that our manipulation left a wide enough margin until the first 30-minute interval was reached. Consequently, there should be no differences in the working time charged across the treatments specified below. Two other features of our manipulation are also noteworthy. First, except for the manipulation all computers were in perfect shape. Hence, any kind of additional repair or service constitutes overtreatment. Second, the costs for a proper repair only include working time, since no spare parts are necessary, and are thus rather low. Specifically, our IT department estimated an average repair price of 30 to 50 Euro. This means that it makes sense to perform the repair – a feature that would not be fulfilled if the diagnosis and the repair were very costly relative to the computer's value of about 540 Euro.

The shops for the first wave of our field experiment were selected as follows: We first compiled a list of all repair shops in the respective cities using information available online (Google, Yellow Pages, city directory, etc.) and assigned to each shop a specific number. Then, we used a random number generator (implemented in Wolfram's Mathematica) to re-rank the shops. Beginning with the lowest rank on this re-ranked list, we then randomly allocated one of our treatments (as specified below) to each shop making sure that we have basically the same number of observations for each of the treatments.

The interaction with the computer repair shops was implemented in a double-blind fashion in the following way. First, we wrote an email (from a private email address using a signature containing a postal address in the respective city) to a repair shop with the following fixed text: *"Hi! I dropped my laptop and now it is no longer able to boot. I only get a black screen and some beep signals. I wanted to ask if I can bring the laptop in for repair."* After a repair shop

[11] See https://support.lenovo.com/at/en/solutions/ht035729 (accessed on 10 January 2019).

had confirmed that we could bring the laptop we sent the actual treatment variation again via mail. Only after this treatment variation, we sent experimental helpers with the computer to the repair shop. The helpers were unaware of our research question, the treatment variations and of course the treatment to which a specific shop had been assigned to. Using such a double-blind procedure minimizes the interaction between helpers and repair shop staff and implies that the personal encounter when dropping off the computer has no direct impact on the size of the treatment effects.[12]

The following three treatments were implemented and led to the predictions described in the following:

- BASELINE: In this treatment, the second email to the shop read as follows: "*Hi! Thanks for your response. A friend of mine will drop off the laptop in the course of this week. The password of the laptop is: "veronika123". Please inform me as soon as you know more.*" Here we did not mention any potential source of the problem, since this reflects a typical case in a credence goods market, given that consumers are usually not aware of the prevailing problem and their associated needs.
- CORRECT: In this treatment, we started with the identical script as in BASELINE, but then added the following text: "*I informed myself a bit on the internet and I think that the beep is caused by a problem of the RAM modules. Maybe this helps.*" In this case the owner's self-diagnosis is correct. In theory, repair services are credence goods for consumers who are unable to self-diagnose the problem, but ordinary goods for other consumers (see Dulleck and Kerschbamer, 2006). At first sight, one might therefore expect lower repair prices in CORRECT. However, recall that the diagnosis is still noisy, and other sources might cause the problem. For the latter reason, it is unclear what to expect from a comparison of prices in BASELINE and CORRECT.
- INCORRECT: Here, we started with the identical script as in BASELINE, but then added the following text: "*I informed myself a bit on the internet and I think that the beep is caused by a problem of the motherboard. Maybe this helps.*" In this case the owner's self-diagnosis is wrong. Yet, recall that the acoustic error message suggested a problem with the main board as a potential, although less likely, source of the problem. This means that our script conveys a realistic possibility, and it does not signal incompetence on the side of the consumer (which shops might be tempted to exploit). It is important to note that our treatment variation was deliberately designed

[12] Our helpers were instructed simply to drop off the computer at the repair shop and keep the interaction as short as possible.

in such a way that the wrong self-diagnosis by the customer is easily detected by a competent repairer. Indeed, our IT department assured us that a competent shop would immediately notice that the motherboard did not cause the problem and that the latter was due rather to the loose RAM modules. So, even if a repair shop took the incorrect self-diagnosis of the consumer as the starting point, generating a correct diagnosis and repairing the PC should be easily done within 30 minutes, meaning that the price an honest shop charges in INCORRECT should not be higher than the price charged in BASELINE. However, the incorrect self-diagnosis arguably generates more room for a dishonest expert seller to overtreat or overcharge the customer. Assuming that some shops exploit this opportunity, we expected higher prices in total in INCORRECT than both in BASELINE and CORRECT.

This first wave of our experiment was conducted between November 2015 and July 2016. We sent seven undercover helpers ("mystery shoppers") with one manipulated computer each during regular opening hours to the repair shops on our list. As specified above, treatment assignment was random and experimental helpers were blind to the treatment.

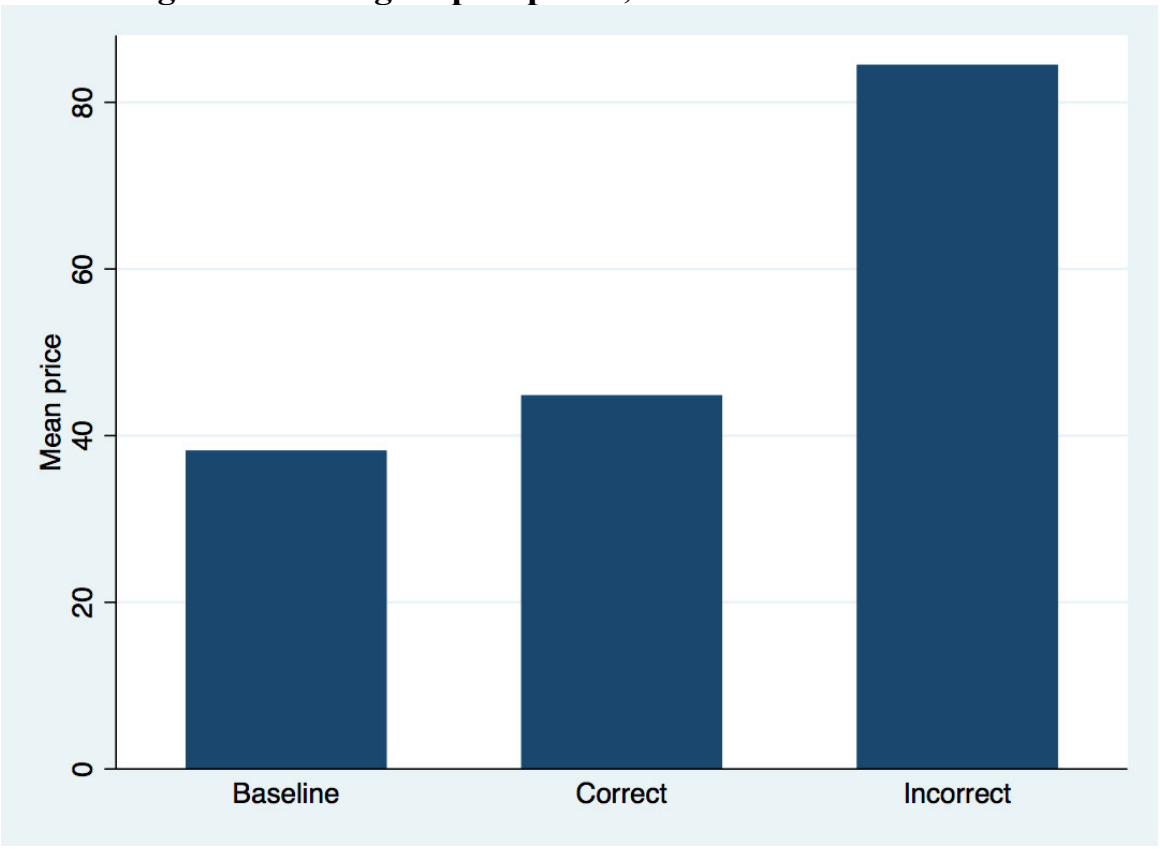
3. Results of Wave 1 – The effects of a self-diagnosis from the web

In total, we collected 71 observations; 24 for BASELINE, 24 for CORRECT, and 23 for INCORRECT. Out of these shops, one shop in INCORRECT claimed that the computer could not be repaired and charged 20 Euro for the diagnosis. We exclude the latter shop from the data analysis, since no service was provided in this case. This leaves us with 70 observations for the analysis.

The mean duration from handing in the computer to getting it back was 3.53 days, and the median was 2 days (with practically no differences across treatments). When we picked up the computer, it was not always possible to identify from the bill – or even upon request in the shop if the working time was not shown on the bill – how much working time was charged for the repair.[13] From all cases with known working time (50 of the 70 shops in our data base), we see that the average working time was shortest in BASELINE (0.51 hours, N=20), intermediate in CORRECT (0.62 hours, N=16) and longest in INCORRECT (0.86 hours, N=14).

[13] According to German law, it is not mandatory to indicate the working time on the bill.

Figure 1. Average repair prices, conditional on treatment



Given that there are no significant treatment differences in the likelihood of a successful repair, the repair price – which is obviously observable in all cases – is the key indicator of service provision.[14] Figure 1 shows the average price across the three treatments. It is 38.21 Euro in BASELINE (N=24), 44.85 Euro in CORRECT (N=24), and 84.50 Euro in INCORRECT (N=22). Figure 2 presents the cumulative distribution function, showing that the variance of repair prices is considerably larger in INCORRECT than in the other two treatments.

[14] From the 70 observations in our data base, one shop saved the data on an external hard drive, but did not repair the computer. All the other shops were able to repair the laptop.

Figure 2. Cumulative distribution function of repair prices, conditional on treatment

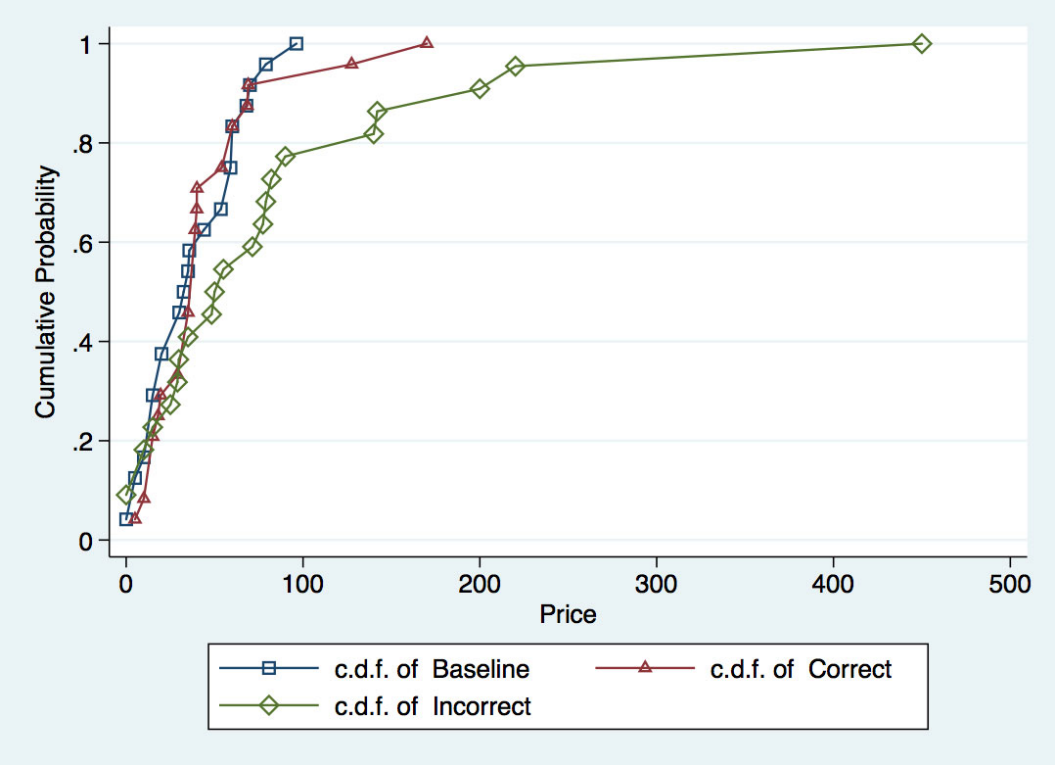


Table 1 presents an OLS-regression with the repair price as the dependent variable and two dummies for the treatments CORRECT and INCORRECT, as well as a number of control variables as independent variables. These variables are motivated as follows: The idea behind controlling for whether a shop is run by a single person or not ("One-man business") is that, in the case of a one-man business, the owner is the single residual claimant of all revenues. This means that one-man businesses might have less diluted incentives to charge higher prices, compared to multi-person shops where employees typically receive fixed wages. Additionally, inspired by Mimra et al. (2016) and Rasch and Waibel (2018), we control for the "number of competitors"– i.e., the number of other repair shops within a circle of 5 kilometers – to account for possible competition effects. Finally, we control for the average "rental price" of apartments in the district.[15] This variable may be important for two reasons. First, it can be taken as a proxy for the average wealth in a given district from which typically customers are attracted. If expert sellers engage in price discrimination of customers (Gneezy et al., 2012), the average rental price may turn out to have a positive coefficient. Second, the rental price may capture an

[15] This information was taken from www.wohnungsboerse.net.

important element of a shop's cost function, namely the rent for the shop, which is why shops in more expensive districts might charge higher prices.

Table 1. Regression analysis of repair prices

Dependent variable (OLS regressions) Independent variables	*[1]* *Repair price (in Euro)*	*[2]* *Repair Price (in Euro)*
CORRECT treatment (1=yes)	16.09	26.51*
	(10.78)	(13.51)
INCORRECT treatment (1=yes)	53.56**	62.07**
	(23.42)	(25.67)
One-man business (1=yes)	15.37	28.80*
	(14.10)	(15.67)
Number of competitors within 5 km	1.66**	1.96**
	(0.76)	(0.87)
Rental price in the district of the shop (€/m²)	1.72	0.01
	(1.55)	(1.44)
Constant	-10.50	-15.47
	(23.41)	(28.27)
Negative ratings (log of number of ratings with 1 star or 2 stars plus 1)		30.04**
		(15.01)
Positive ratings (log of number of ratings with 3 stars or better plus 1)		-1.66
		(4.56)
# Observations	70	70

OLS-regressions (robust standard errors) with repair price (in Euro) as independent variable, including, as explanatory variables, a treatment dummy for CORRECT, a treatment dummy for INCORRECT, a dummy for being a one-man business, the number of other shops within a radius of 5 km, and an index for the average rental price in the district of the shop.

***, **, * denote significance at the 1%, 5%, 10% level, standard errors in parentheses.

Column [1] of Table 1 presents the results of our OLS-regression. We take BASELINE as the benchmark, and find that CORRECT increases prices, but not significantly so. This at least means that a correct self-diagnosis by the customer does *not reduce* repair prices compared to the BASELINE where consumers do not mention any possible source, but simply describe the problem when handing in the computer for repair. Contrary to the null-effect of a *correct* self-

diagnosis, however, the treatment effect of INCORRECT is economically very large and statistically significant. Controlling for the other explanatory variables, the estimated price difference between BASELINE and INCORRECT is almost 50 Euro. Given an average repair price of 38 Euro in BASELINE, this implies an estimated price increase of about 130% in case the consumer states a wrong diagnosis. Recall that our IT department predicted the wrong diagnosis in treatment INCORRECT ("problem with the motherboard") to be quickly recognized, and that finding out about the real RAM problem and fixing it should be possible in less than 30 minutes, which is the first unit of working time that shops charge for.

Our main treatment results are robust to controlling for the additional explanatory variables described above and included in column [1]. The dummy for the number of competitors within 5 km is significantly positive, confirming earlier results of Dulleck et al. (2011) and Mimra et al. (2016), namely that competition increases the likelihood of fraudulent behavior on the seller side. One explanation for the price-increasing effect of competition might be that shops with a larger number of competitors are typically more centrally located in the cities under investigation. A more central location might mean more occasional customers who can be exploited more easily without being punished by the market in the future. A more decentralized shop – with fewer competitors around – may have more regular customers with repeat visits and hence may be inclined to charge them less (similar to findings in Schneider, 2012). The dummy for a one-man business is positive, but insignificant in column [1], and so is the dummy for the average rental price in the district of the shop. We summarize our main findings from Table 1 as our first result and our first implication:

Result 1 and Implication 1: *A correct self-diagnosis does not reduce prices, while an incorrect one increases repair prices substantially. Since the diagnosis provided by specialized software on the internet is almost always noisy and since consumers cannot distinguish between a correct and an incorrect diagnosis, an implication of our result is that acquiring and revealing a noisy self-diagnosis is, in expectation, a costly mistake for consumers in markets for credence goods.*

So far, we have analyzed whether it is a good strategy for consumers to use information provided on the internet for a (possibly noisy) self-diagnosis of their needs (in our case: the source of the computer problem). As Result 1 shows, revealing such a self-diagnosis to the expert seller is a costly mistake. After having concluded this wave of our field experiment, we realized that we could look *ex post* into the internet ratings of the repair shops in our database.

Doing so would allow us to identify whether rating platforms contain useful information for consumers even in the case of credence goods markets in which consumers typically cannot judge their exact needs even after service or goods provision by sellers.

We looked up the internet ratings of our 70 shops on Yelp and Google (because these platforms have the most reviews about repair services) and classified them into positive and negative ratings. As negative ratings, we took those with 1 or 2 stars, and as positive ratings those with 3, 4, or 5 stars. In Appendix A we show in Table A1 that the results do not change qualitatively if we would classified all ratings with 1, 2, or 3 stars as negative, and those with 4 or 5 stars as positive.

Column [2] of Table 1 adds the logs of the number of positive and negative ratings (adding 1 to this number in order not to lose shops without any of these ratings) to the variables already included in column [1]. We see that negative ratings have a significantly price-increasing effect. Positive ratings, however, are insignificant (but have a negative sign, as expected). Adding the ratings changes a few of the other results from column [1]. First, the dummy for CORRECT is now weakly significantly positive, indicating that revealing a self-diagnosis is a costly mistake, not only in expectation but even for the (best) case where the self-diagnosis is correct. Second, the dummy for a one-man business becomes weakly significantly positive in column [2], supporting the hypothesis that if shop owners are the full residual claimants they have higher incentives to charge higher prices (for the same service).

4. The usefulness of internet ratings – Wave 2 of the field experiment

Given that in wave 1 of our experiment we accessed and considered internet ratings only *ex post* after collecting the data on repair prices, we added a second wave of data collection to analyze the usefulness – but also the potential pitfalls – of internet ratings in credence goods markets in more detail and with an *ex ante* hypothesis.

Wave 2 was run in March and April 2017. We used the same RAM-manipulation as described above for the first wave. This time, however, we did not implement different treatments, but had only what we call a BASELINE-2 condition. Since we had no treatment variations, double-blindness was no longer an issue. For this reason, the communication with the repair shops was no longer via e-mail. Instead, the mystery shoppers (i.e., our experimental helpers) approached the shops directly with the manipulated computer and the following script: *"Hi! I dropped my laptop and now it is no longer able to boot. I only get a black screen and some beep signals. I wanted to ask if you can repair it."*

Data were collected in a new city, Berlin, the largest city in Germany. Based on power calculations that we derived from the first wave, we aimed at collecting 60 observations to detect a price difference between better-rated shops and worse-rated shops at a 5%-significance level with 80% power.[16] To assess the empirical relation between ratings and repair prices we decided to include in our sample those computer repair shops in Berlin that had the largest number of internet reviews on Yelp and Google. Of the more than 100 repair shops in this new city, 58 shops had 3 or more reviews. Those 58 shops were targeted in wave 2 of our experiment. We sent our experimental helpers to each of these shops with a request for a repair. Unknown to the helpers, we started with the hypothesis – based on our wave 1 results – that those shops whose average rating is worse (i.e., lower) than the median average rating across all shops would charge higher prices than the shops whose average rating is better (i.e., higher) than this median.

5. Results of Wave 2 – The predictive power of internet ratings and the problem with non-recommended reviews

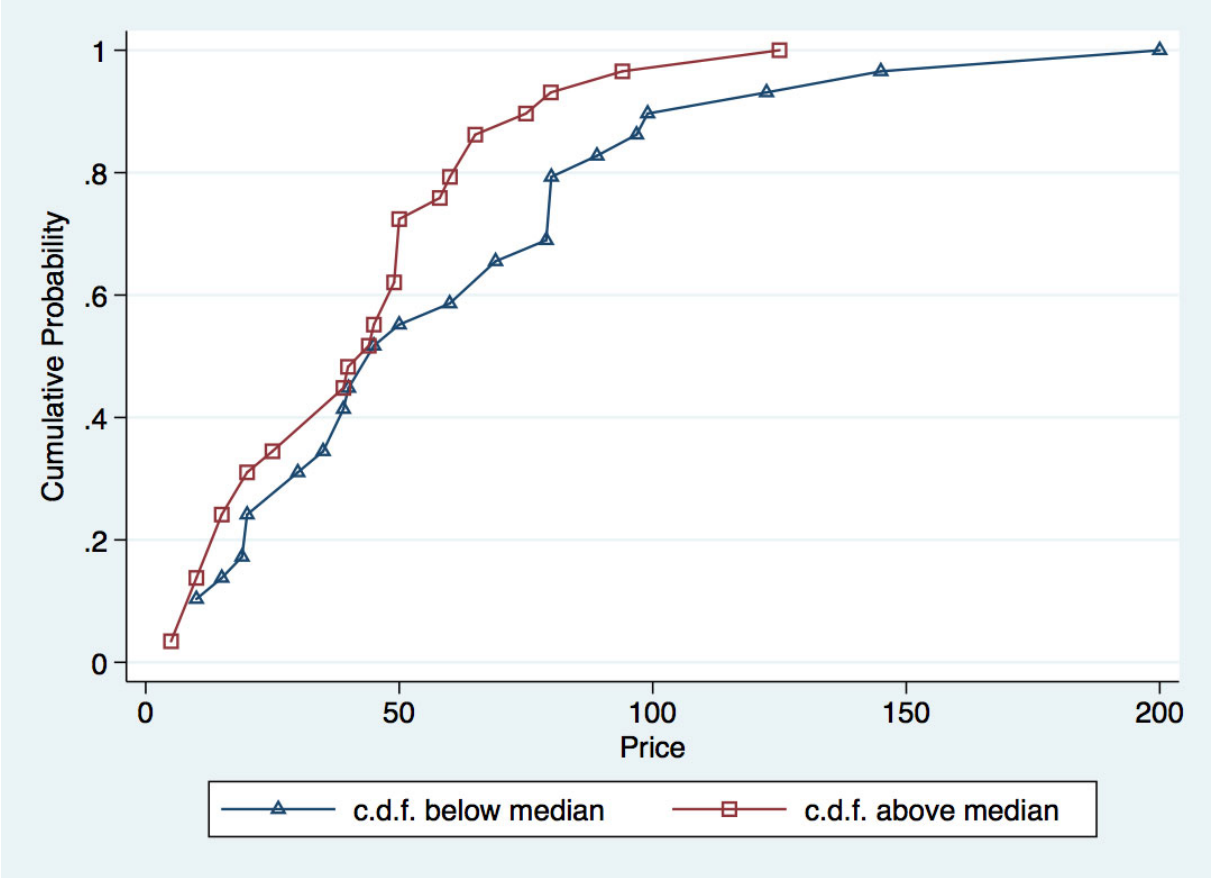
We first look at the repair prices of shops, contingent on their average rating. For that we perform a median split of the average rating of all 58 shops. Those shops with an average rating below the median charged on average 59.52 Euro (N=29), while those with an average rating above the median had a significantly lower average price of 43.48 Euro (N=29), which confirms our directional hypothesis ($p < 0.05$; one-sided t-test). Figure 3 shows the cumulative distribution function of prices, with the graph for the shops rated below the median lying always to the right of the better-rated shops. Recall that all 58 shops were handling completely identical (and identically manipulated) computers. Since all shops were able to repair the computer, it is striking – to our mind – that the internet reviews are a significant predictor of which set of shops charges higher prices for the same (successful) repair than others. This is summarized in our second result.

Result 2 and Implication 2: *Internet ratings are indicative of repair prices: Shops with an average rating below the median charge on average significantly higher prices than shops*

[16] For our power calculations, we considered the observations in BASELINE in the first wave as our basis for the effect size of ratings, performed a median split of the average rating of all shops with at least two reviews, calculated the average price for above-median (25.84€) and below-median shops (47.61€), and then performed the calculation to get a difference between above-median and below-median shops at the 5%-significance level with 80% power.

with an average rating above the median. An immediate implication is that consumers can profit from the internet ratings of former consumers – even in markets where informational asymmetries continue to be present even after consumption.

Figure 3. Cumulative distribution function of repair prices, conditional on average rating above or below median



Result 2 is encouraging, but there might be more to ratings than just looking at averages and consider all ratings as equally valuable. After all, sellers have incentives to get good reviews and ratings, as good reviews may attract consumers and thus increase revenues and profits. Good customer service that pays off in customers writing nice reviews is one means for increasing the number of positive ratings. An alternative means for shop owners to increase their average rating is to "order" good reviews from friends or even fake good reviews through self-generated user profiles. There is abundant evidence that internet review platforms are not immune to this type of fraudulent behavior on the side of sellers (see, e.g., Ockenfels and Resnick, 2012; Streitfeld 2012; Luca and Zervas, 2016).

Given the possibility that some reviews might be manipulated or even faked, it would potentially be useful to discriminate between more and less reliable reviews in order to see

whether those two types of reviews have different predictive value for the repair prices of our manipulated computers. Actually, a specific feature of the review platform Yelp allows one to distinguish between recommended and non-recommended reviews, and in the following we exploit this feature to examine whether these two types of reviews differ in their predictive power for prices.

On its internet site, Yelp explains its classification in recommended, respectively non-recommended, reviews as follows: "We use automated software to recommend the reviews we think will be the most helpful to the Yelp community based primarily on quality, reliability, and the reviewer's activity on Yelp."[17] This means that Yelp tries to filter out fake reviews, for instance where shops are supposed to write about themselves, or reviews from reviewers who have a poor reputation in the community.

In Table 2, we present, for the 58 shops in our database, the average number of recommended and non-recommended reviews and the distribution of ratings from 1 star to 5 stars.[18] Column [1] shows data for recommended reviews, and column [2] for non-recommended ones. The first thing to note is that only 23% of the total number of reviews are recommended by Yelp. A second important observation is that the average rating of recommended reviews is (at 3.94) lower than the average rating of non-recommended ratings (4.28). The difference is highly significant (p=0.0017, two-sided t-test), indicating that non-recommended reviews are systematically more positive than recommended ones. This is mainly due to the much larger fraction of 5-star ratings in the set of non-recommended reviews than in the recommended reviews. In column [3] of Table 2, we show the reviews from Google for the 58 shops in our database. Google does not offer a distinction between recommended and non-recommended shops. For this reason, we are going to treat the Google reviews as separate from the two sets of Yelp reviews in the following analysis, in which we address the question whether the three data sets are related in different ways to repair prices in wave 2 of our field experiment.

[17] See the "Not Recommended" section on the following webpage: https://www.yelp.com/biz/notebookservice030-berlin-3?osq=laptop+reparatur (accessed on 10 January 2019).
[18] These data were retrieved from the rating platforms in June 2017.

Table 2. Descriptive statistics of ratings on Yelp and Google (Wave 2)

	[1] YELP RECOMMENDED (N=241 reviews)			[2] YELP NON-RECOMMENDED (N=825 reviews)			[3] GOOGLE (N=1,518 reviews)		
	Mean	Min.	Max.	Mean	Min.	Max.	Mean	Min.	Max.
Number of reviews per shop	3.69 (5.02)	0	27	14.22 (30.18)	0	145	26.17 (33.10)	3	227
Mean rating per shop	3.94 (1.46)			4.28 (1.43)			4.25 (1.44)		
Number of 1 star ratings	0.57 (0.57)	0	5	2.02 (5.05)	0	33	3.66 (4.87)	0	27
Number of 2 star ratings	0.12 (0.42)	0	2	0.26 (0.95)	0	5	0.71 (1.33)	0	7
Number of 3 star ratings	0.28 (0.81)	0	5	0.24 (0.66)	0	4	0.59 (1.33)	0	8
Number of 4 star ratings	0.74 (1.46)	0	6	0.90 (2.61)	0	17	1.72 (4.50)	0	32
Number of 5 star ratings	1.98 (2.93)	0	15	10.79 (23.72)	0	109	19.5 (24.40)	0	153

Standard deviation in parentheses.

In Table 3, we regress the repair prices in wave 2 on positive and negative ratings and on the three control variables that we had already used and described in Table 1. With regard to the ratings, we distinguish between recommended and non-recommended reviews on Yelp, keeping Google reviews as a separate category.[19]

[19] In Table A2 in the appendix we show an alternative specification where all ratings with 1, 2, or 3 stars are classified as negative, while those with 4 or 5 stars are classified as positive. The qualitative results remain unchanged.

Table 3. Recommended and non-recommended reviews and repair price

Dependent variable (OLS-regressions)	*Repair price*
Independent variables	*(in Euro)*
Recommended reviews on Yelp	
Negative ratings (log number of 1-star & 2-star ratings plus 1)	23.85**
	(10.24)
Positive ratings (log number of 3-star, 4-star, and 5-star ratings plus 1)	-17.00**
	(7.81)
Non-recommended reviews on Yelp	
Negative ratings (log number of 1-star & 2-star ratings plus 1)	2.33
	(7.72)
Positive ratings (log number of 3-star, 4-star, and 5-star ratings plus 1)	15.12**
	(6.63)
Reviews on Google	
Negative ratings (log number of 1-star & 2-star ratings plus 1)	-5.71
	(6.58)
Positive ratings (log number of 3-star, 4-star, and 5-star ratings plus 1)	-13.63**
	(5.96)
One-man business (1=yes)	11.46
	(10.92)
# Competitors within 5km	-1.04
	(1.23)
Rental price in the district (€/m^2)	-0.97
	(3.21)
Constant	99.64***
	(36.02)
# Observations (repair shops)	58

OLS-regressions with repair price (in Euro) as independent variable, including, as explanatory variables, positive and negative ratings of recommended andnon-recommended reviews on Yelp, and on Google (where there is no distinction between recommended and non-recommended reviews). Additional controls like in Table 3 apply. ***, **,* denote significance at the 1%, 5% and 10% level, standard errors in parentheses.

The first two explanatory variables in Table 3 draw on recommended reviews on Yelp and they reveal a pattern that matches our *ex ante* hypothesis: Negative ratings are associated with significantly higher repair prices, while positive ratings correlate significantly with lower prices. This confirms the result that we had obtained *ex post* from wave 1 in the *ex ante* setting of wave 2. Yet, we are now able to examine in a more refined way how ratings are related to actual repair prices. The second set of explanatory variables refers to non-recommended reviews, and here we note that negative ratings are insignificant, while positive ratings have a significantly *positive* effect on prices – which is exactly opposite to the effect of recommended positive reviews on Yelp! We consider the latter a striking, albeit not entirely unexpected, finding. In fact, it lends credibility to Yelp's classification of non-recommended reviews, as non-recommended positive reviews seem to misguide consumers. Negative ratings that are non-recommended do not have a significant effect, implying that they do not contain useful information for consumers.[20] Reviews on Google have a price-decreasing effect when they are positive, but no significant effect when they are negative. The latter result is in contrast to the significantly price-increasing effect of recommended negative reviews on Yelp, but in line with the non-significant impact of non-recommended negative reviews on this platform. So, compared to recommended Yelp reviews, reviews on Google seem to have much less informational value. This may be due to the different ways of dealing with fake reviews on Yelp and on Google – with the latter being much more lenient towards companies suspected of producing fake reviews.[21] The three additional controls at the bottom of Table 3 remain insignificant, which is partly different from the findings in Table 1. A potential explanation for the insignificance of these control variables is that in Table 3 we capture more information from the ratings than we were able to do in Table 1 – where we did not differentiate between recommended and non-recommended ratings. Overall, we summarize the main findings and implications from looking deeper into the informational value of internet reviews as our third result as follows:

Result 3 and Implication 3: *The informational value of internet ratings is heterogeneous: the most informative ratings are recommended ones (on Yelp) where negative ratings have*

[20] In principle, non-recommended negative ratings could well have a significant price-decreasing effect. This would be the case, for instance, if shops wrote or commissioned negative reviews for their most dangerous competitors in order to look better themselves in relative terms (see https://www.nytimes.com/2011/05/22/your-money/22haggler.html?_r=2 for casual evidence pointing in this direction). If correctly identified as fake reviews by the Yelp software, this kind of unfair competition could lead to a significant negative correlation between non-recommended negative ratings and repair prices. We do not find such an effect.

[21] See the elaborate discussion of Joy Hawkins on https://searchengineland.com/yelp-vs-google-how-do-they-deal-with-fake-reviews-307332 (accessed on 26 January 2019).

strongly price-increasing effects and positive ratings strongly price-decreasing effects; for non-recommended positive reviews the correlation is exactly reversed (non-recommended positive ratings increase prices), while non-recommended negative reviews are not informative for prices. Together these results suggest that consumers in credence goods markets can benefit from the wisdom of crowds – manifested in internet ratings – but that they should take the distinction between recommended and non-recommended ratings seriously and they should rely more on platforms that have more restrictive filters for potentially commissioned or fake reviews.

6. Conclusion

Modern communication technologies have transformed and often disrupted markets in a significant way. For instance, digital platforms like eBay, Amazon, or Airbnb have expanded the scope of trade by making a match between sellers and buyers much easier than in former (offline) times (Roth and Ockenfels, 2002; Bolton et al., 2013, 2019). Digitization has also affected labor markets by making extremely flexible work contracts with extraordinarily adjustable working hours feasible (like with Uber, where drivers can practically decide themselves when and how long they want to work, thus creating a positive driver value of this flexibility; Chen et al., 2019). Modern communication technologies might also have an important impact on markets for credence goods where informationally disadvantaged consumers are systematically exploited by better informed experts. The size of these markets is huge – and the issue of fraudulent behavior on the seller side looms large (Iizuka, 2007; Schneider, 2012; Kerschbamer and Sutter, 2017). The question whether and to which degree consumers on credence goods markets can benefit from new media is therefore a crucial one and it has not been addressed so far in the literature.

The present study has addressed this question by performing a field experiment in a market for credence goods. The starting point of our project has been the conjecture that the digital era has made it much easier – and much cheaper – for consumers to gather information which can help them to diagnose their needs or to assess the trustworthiness of sellers on credence goods markets – and our main research interest was the causal link between information retrieved from the internet and from social media and the extent to which consumers are cheated upon by sellers on credence goods markets.

Our field experiment has been run in the market for computer repairs. In the first wave of our experiment we have found that consumers make a costly mistake when they acquire a noisy

self-diagnosis from specialized webpages about the problem of their computer *and* reveal this self-diagnosis to the repair shop. Mentioning a correct self-diagnosis about the appropriate repair does not reduce the repair price in comparison to a situation where the computer owner simply asks for a repair, and mentioning an incorrect diagnosis more than doubles the average price.[22] This constitutes our first main result.

While improving a consumer's information about the source of the problem does not reduce exploitation through sellers, improving her information about previous customers' experiences with a particular seller might be potentially useful: In the second wave of our experiment we have found that shops with better internet ratings charge significantly lower prices (for the same and successful repair) than shops with worse ratings. This finding constitutes our second main result and it is particularly remarkable because a defining feature of a credence good is exactly that consumers typically cannot judge even after an interaction with a seller whether they received the good or service that maximized their consumer surplus. As a consequence, it is less than straightforward to assume that internet reviews carry useful information for future consumers in credence goods markets.

A third main result of our field experiment has been that not all reviews are equally useful: Motivated by reports in the media about potentially fake or commissioned internet reviews,[23] we have examined how reviews that are classified on one platform (Yelp) as either recommended or non-recommended are related to actual service provision and repair prices. While for recommended reviews we have found that negative reviews are associated with significantly higher prices, and positive reviews with significantly lower prices, our result with respect to non-recommended reviews is rather striking: If such non-recommended reviews are positive, they are associated with significantly *higher* prices, which is completely contrary to the price-decreasing effect of recommended positive reviews.

Of course, discriminating between trustworthy and non-trustworthy reviews is a challenge. In an interview for the New York Times, a media representative of Yelp, Vince Sollitto, stated that Yelp would not reveal its algorithm for classifying reviews as recommended or not. Yet, Sollitto said "our job is to find and filter out fake reviews. At the same time we let our audience know that this system isn't perfect. Some legitimate content might get filtered and some illegitimate content might sneak through. We're working hard at it. It's a tough one."[24] Our third main result suggests that the algorithm used by Yelp to discriminate between

[22] The finding that a correct self-diagnosis does not reduce the repair price resembles the result in Gottschalk et al. (2018) that a second opinion does not reduce overtreatment.
[23] See, for instance, the *New York Times* article: https://www.nytimes.com/2011/08/20/technology/finding-fake-reviews-online.html?_r=1&ref=todayspaper (accessed on 10 January 2019).
[24] See https://www.nytimes.com/2011/05/22/your-money/22haggler.html (accessed on 10 January 2019).

recommended and non-recommended reviews is helpful in identifying reviews that should not be trusted by consumers. On the contrary, reviews on Google have much less informational value for consumers, possibly because of their more lenient stance towards companies that are suspected of fake reviews. In sum, this implies that consumer protection on credence goods markets – and more generally on all markets with informational asymmetries – might not only depend on regulation from governmental authorities, but it might benefit significantly from private review platforms (such as Yelp) that use artificial intelligence to assess the trustworthiness of reviews posted on such platforms.

We are confident that even more can be gained in the future for consumers by analyzing the content of internet reviews, and not only the rating itself, because such an analysis might reveal more insights into which expert sellers take advantage of consumers and which do not. Our natural field experiment has provided a first causal investigation of how social media platforms can inform and benefit consumers on credence goods markets. We consider it as a fruitful direction for future research to study other markets also – for instance, health care markets or the markets for financial or legal advice – and their conditions under which the information from social media is rather a blessing (by helping consumers to get appropriate quality without cheating) than a curse (by falling prey to fake reviews that mislead consumers and expose them to exploitation).

References

Abeler, J., Nosenzo, D., Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, forthcoming.

Akerlof, G.A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488-500.

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

Balafoutas, L., Kerschbamer, R., Sutter, M. (2017). Second-degree moral hazard in a real-world credence goods market. *Economic Journal* 127, 1-18.

Beck, A., Kerschbamer, R., Qiu, J., Sutter, M. (2013), Shaping beliefs in experimental markets for expert services: Guilt aversion and the impact of promises and money-burning options. *Games and Economic Behavior* 81, 145-164.

Bohnet, I., Huck, S. (2004). Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American Economic Review, Papers and Proceedings* 94, 362-366.

Bolton, G., Katok, E., Ockenfels, A. (2004). How effective are electronic reputation mechanisms? An experimental investigation. *Management Science* 50, 1587-1602.

Bolton, G., Greiner, B., Ockenfels, A. (2013). Engineering trust – Reciprocity in the production of reputation information. *Management Science* 59, 265-285.

Bolton, G., Greiner, B., Ockenfels, A. (2019). Dispute resolution or escalation? The strategic gaming of feedback withdrawal options in online markets *Management Science*, forthcoming.

Cappelen, A., Halvorsen, T., Sørensen, E., Tungodden, B. (2017). Face-saving or fair minded: What motivates moral behavior. *Journal of the European Economic Association* 15, 540-557.

Chen, M. K., Chevalier, J. A., Rossi, P. E., Oehlsen, E. (2019). The value of flexible work: Evidence from Uber drivers. *Journal of Political Economy*, forthcoming.

D'Acunto, F., Prabhala, N., Rossi, A. G. (2018). The promises and pitfalls of robo-advising. *Review of Financial Studies*, forthcoming.

Darby, M., Karni, E. (1973). Free competition and the optimal amount of fraud, *Journal of Law and Economics* 16, 67-88.

Dulleck, U., Kerschbamer, R. (2006). On doctors, mechanics and computer specialists – The economics of credence goods. *Journal of Economic Literature* 44, 5-42.

Dulleck, U., Kerschbamer, R., Sutter, M. (2011). The economics of credence goods: An experiment on the role of liability, verifiability, reputation, and competition. *American Economic Review* 101, 526- 555.

Gneezy, U. (2005). Deception: the role of consequences. *American Economic Review* 95: 384-394.

Gneezy, U., Kajackaite, A., Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review* 108(2), 419-453.

Gneezy, U., List, J.A., Price, M.K. (2012). Toward an understanding of why people discriminate: Evidence from a series of natural field experiments. *NBER Working Paper* 17855.

Gottschalk, F., Mimra, W., Waibel, C. (2018). Health services as credence goods: A field experiment. Available at SSRN: https://ssrn.com/abstract=3036573.

Grosskopf, B., Sarin, R. (2010). Is reputation good or bad? An experiment. Am*erican Economic Review* 100(5), 2187-2204.

Huck, S., Luenser, G., Spitzer, F., Tyran, J.R. (2016a). Medical insurance and free choice of physician shape patient overtreatment. A laboratory experiment. *Journal of Economic Behavior and Organization* 131, 78-105.

Huck, S., Luenser, G., Tyran, J.R. (2016b). Price competition and reputation in markets for experience goods. An experimental study. *RAND Journal of Economics* 47, 99-117.

Huck, S., Lünser, G., Tyran, J.–R. (2012). Competition fosters trust. *Games and Economic Behavior* 76(1), 195-209.

Iizuka, T. (2007). Experts' agency problems: Evidence from the prescription drug market in Japan. *The RAND Journal of Economics* 38, 844-862.

Kerschbamer, R., Sutter, M. (2017). The economics of credence goods – A survey of recent lab and field experiments. *CESifo Economic Studies* 63, 1-23.

List, J., I. Rasul (2011). Field experiments in labor economics. In: O. Ashenfelter, D. Card (eds.) *Handbook of Labor Economics*, 4A, 103-228. Amsterdam: Elsevier.

Luca, M., Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science* 62, 3412-3427.

Mimra, W., Rasch, A., Waibel, C. (2016). Price competition and reputation in credence goods markets: Experimental Evidence. *Games and Economic Behavior* 100, 337-352.

Ockenfels, A, Resnick, P. (2012). Negotiating reputations. In: G. Bolton, R. Croson (eds), *The Oxford Handbook of Economic Conflict Resolution,* Oxford University Press, 223-237.

Rasch, A., Waibel, C. (2018). What drives fraud in a credence goods market? Evidence from a field study. *Oxford Bulletin of Economics and Statistics* 80, 605-624.

Roth, A., Ockenfels, A. (2002). Last minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the internet. *American Economic Review* 92, 1093-1103.

Schneider, H. S. (2012). Agency problems and reputation in expert services: Evidence from auto repair. *Journal of Industrial Economics* 60, 406-433.

Streitfeld, D. (2012), The best book reviews money can buy. http://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html?pagewanted=1&_r=2&partner=rss&emc=rss〉. Accessed on 10 January 2019.

Zitzewitz, E. (2012). Forensic economics. *Journal of Economic Literature* 50, 731-769.

Appendix

A. Additional tables

Table A1. Regression analysis of repair prices

	[1]
Dependent variable (OLS regressions)	*Repair Price*
Independent variables	*(in Euro)*
CORRECT treatment (1=yes)	23.59*
	(12.21)
INCORRECT treatment (1=yes)	60.49**
	(25.30)
One-man business (1=yes)	25.12*
	(15.06)
Number of competitors within 5 km	1.94**
	(0.88)
Rental price in the district of the shop (€/m²)	0.25
	(1.45)
Constant	-14.87
	(28.67)
Negative ratings (log of number of ratings	20.72**
with 1 star, 2 stars, or 3 stars plus 1)	(9.49)
Positive ratings (log of number of ratings	-0.80
with 4 stars or better plus 1)	(6.38)
# Observations	70

OLS-regressions (robust standard errors) with repair price (in Euro) as independent variable, including as explanatory variables a treatment dummy for CORRECT, a treatment dummy for INCORRECT, a dummy for being a one-man business, the number of other shops within a radius of 5km and the rental prices in the district of the shop.

***, **, * denote significance at the 1%, 5%, 10% level, standard errors in parentheses.

Table A2. Recommended and non-recommended reviews and repair price

Dependent variable (OLS-regressions)	*Repair price (in Euro)*
Independent variables	
Recommended reviews on Yelp	
Negative ratings (log number of 1, 2 & 3 star ratings plus 1)	16.23*
	(9.13)
Positive ratings (log number of 4 and 5 stars ratings plus 1)	-16.76*
	(8.95)
Non-recommended reviews on Yelp	
Negative ratings (log number of 1, 2 & 3 star ratings plus 1)	-3.69
	(8.66)
Positive ratings (log number of 4 and 5 stars ratings plus 1)	18.73**
	(7.27)
Reviews on Google	
Negative ratings (log number of 1, 2 & 3 star ratings plus 1)	-6.30
	(7.13)
Positive ratings (log number of 4 and 5 stars ratings plus 1)	-13.07**
	(5.73)
One-man business (1=yes)	11.07
	(12.02)
# Competitors within 5km	-0.83
	(1.43)
Rental price in the district	-1.48
	(3.50)
Constant	102.79***
	(38.24)
# Observations	58

OLS-regressions with repair price (in Euro) as independent variable, including as explanatory variables positive and negative ratings of recommended, respectively non-recommended, reviews on Yelp, and on Google (where there is no distinction between recommended and non-recommended reviews). Additional controls like in Table 3 apply.

***, **,* denote significance at the 1%, 5% and 10% level, standard errors in parentheses.

B. Specification of notebooks

We bought identical notebooks to be able to speed up the data collection process by bringing each of them to a different shop. The notebooks were completely refurbished and the cost was about EUR 540 for each. The notebooks had the following configuration: Lenovo ThinkPad X230 – 2325-B15 refurbished/ Intel Core i5-3320 Processor 2 x 40 GHz/ 8192MB DDR3/ 128-GB SSD hard disk drive Samsung 850 Pro/ 1,366 × 768 pixel (HD) flat display/Intel HD Graphics 4000/ Intel 6300 AGN Wireless/ Windows 7 Professional (64 bit)/ 12-mo warranty/new battery.