Journal of Public Economics 222 (2023) 104891

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Credence goods markets, online information and repair prices: A natural field experiment ☆

Rudolf Kerschbamer [a], Daniel Neururer [b], Matthias Sutter [c,d,a,*]

[a] University of Innsbruck, Austria
[b] University of Otago, New Zealand
[c] Max Planck Institute for Research on Collective Goods Bonn, Germany
[d] University of Cologne, Germany

ARTICLE INFO

Article history:
Received 24 May 2022
Revised 6 March 2023
Accepted 9 April 2023

JEL-Code:
C93
D82

Keywords:
Credence goods
Fraud
Online information
Rating platforms
Field experiment

ABSTRACT

Credence goods markets are characterized by large informational asymmetries between consumers and expert sellers. In two waves of a natural field experiment in the market for computer repairs we study whether consumers benefit from accessing online information about their needs or previous consumers' experience with particular sellers. We find that gaining noisy knowledge about one's needs and revealing it to the seller is a costly mistake, since seemingly better informed customers pay, on average, higher prices. By contrast, accessing online ratings helps identifying sellers who provide appropriate quality at reasonable prices, in particular on rating platforms that filter out untrustworthy reviews.

1. Introduction

Markets for credence goods (Darby and Karni, 1973; Dulleck and Kerschbamer, 2006; Huck et al., 2016a; Gottschalk et al., 2020) 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. Combined, these markets have a huge size in the overall economy.[1] Their key feature is the informational asymmetry between expert sellers and consumers: Doctors, mechanics, and legal or financial advisors are typically much better informed than patients, clients or private investors about the quality of a good, service or asset that fits a consumer's needs best. 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

☆ We thank the co-editor, Robert Metcalfe, and two anonymous referees for their valuable comments and suggestions. Thanks are also due to Loukas Balafoutas, Alexander Cappelen, Niall Flynn, Ben Greiner, Axel Ockenfels, Henry Schneider, Marco Schwarz, Bertil Tungodden, 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 for helpful comments. Brian Cooper helped editing the document. Financial support from the Austrian Science Fund (FWF) through special research area grants SFB F06305 and SFB F06306, as well as through Grant No P26901 and P27912, and from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanýs Excellence Strategy – EXC 2126/1– 390838866 is gratefully acknowledged.

* Corresponding author at: Max Planck Institute for Research on Collective Goods Bonn, Germany.

E-mail addresses: matthias.sutter@uibk.ac.at, matthias.sutter@coll.mpg.de (M. Sutter).

[1] For instance, health care expenditures alone account for about 10% of GDP in a group of 16 OECD-countries (https://www.oecd.org/els/health-systems/health-expenditure.htm); in the U.S.A., the finance sector accounts for about 8% of GDP (see https://apps.bea.gov/iTable/iTable.cfm?reqid=150&step=2&isuri=1&categories=gdpxind); and computer repair services generate about 999 million Euro per annum in Germany alone (https://www.destatis.de/EN/Themes/Economic-Sectors-Enterprises/Enterprises/ICT-Enterprises-ICT-Sector/Tables/ictb-03-enterprises-persons-employed-turnover-investments.html). Links accessed on 14 December 2022.

[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 typically not possible with credence goods.

https://doi.org/10.1016/j.jpubeco.2023.104891
0047-2727/

expert sellers to cheat on consumers, particularly through overprovision or overcharging (Dulleck and Kerschbamer, 2006). 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. In the short run, overcharging is a pure transfer from the consumer to the expert. In the long run, it might also lead to inefficiencies if the fear of getting overcharged deters consumers from trading on such markets in the future (like in Akerlof's, 1970, analysis of lemons markets), or induces them to search for second opinions.

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 of 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 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 and to what extent the information available on the internet actually helps consumers.

In this paper, we present a novel natural field experiment in the computer repair market to examine the causal effects of information retrieved from the internet and social media on the provision of credence goods.[3] We investigate two main channels that can help consumers contain the degree with which sellers exploit their informational advantages. The first channel works through specialized internet pages that allow consumers to self-diagnose their needs, thereby reducing the extent of the informational asymmetry.[4] The second channel refers to internet ratings of previous consumers. This source of information can help to identify expert sellers who provide appropriate quality at reasonable prices, which could then limit overcharging and overprovision and thus improve efficiency.

There are numerous examples for the first channel. In markets for health care services, for example, some webpages allow patients to enter their symptoms and then generate a diagnosis. In other cases, patients can even upload X-rays to internet portals to get an opinion about their health problems.[5] Smart-phone applications like Google Maps have made it very easy – and practically costless – for taxi 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 credence goods markets (Balafoutas et al., 2013, 2017). As a final example, several webpages allow consumers to self-diagnose the problem in case a computer can no longer be booted.[6] We are going to exploit the latter source of information for consumers in our field experiment.

The second channel through which modern technologies might help consumers on credence goods markets is the plethora of rating platforms (like Yelp or on Google) on which consumers give feedback and rate sellers of different types of goods and services. Some of these platforms refer to credence goods providers, such as physicians, repair shops, or lawyers.[7] With regards to taxi drivers, ratings of these credence goods providers are already inbuilt in Uber's services, for example, as a quality control measure. If reliable, the information contained on rating platforms might help consumers by guiding them to trustworthy expert sellers. This could potentially increase the trade volume and efficiency on credence goods markets.

So far, there have been only few attempts to measure whether modern technologies actually help consumers to receive appropriate provision of credence goods and to get overcharged less than when these technologies are not used. Referring to the first channel of information for consumers discussed above (searching information about one's needs in the internet and revealing it to an expert), the literature most closely related is the one investigating to what extent gathering second opinions from other experts could help consumers in credence goods markets. Gottschalk et al. (2020) conduct a natural field experiment in the market for dental care and in one treatment the undercover patient indicates to the dentist that he has uploaded his dental X-ray to an internet platform where dentists offer free advice, and that he is awaiting a response. The authors find no treatment difference in this respect, with overtreatment rates decreasing only slightly and insignificantly in this condition. The main difference to our design is that our consumers reveal specific information (retrieved from the internet) whereas in Gottschalk et al. (2020) consumers only mention that they will get a second opinion in the future. Bindra et al. (2021) show in the German market for computer repairs that mentioning explicitly that another expert has already been visited neither increase the rate of successful repairs nor decrease the average repair price charged by sellers.[8] A potential explanation for the ineffectiveness of revealed second opinions could be that expert sellers infer from the script that the consumer will most likely accept the next recommendation because every expert visit is associated with search (and in most cases also diagnosis) costs and because a third visit is less likely to reduce the repair price than a second visit. If this was indeed the case, expert sellers have no incentives to lower their prices when facing a revealed second opinion.

In this paper we follow a different approach and study the effects of the much cheaper collection of information through the

[3] Seen from a broader perspective, our paper relates to the literature on what drives individual propensities to act morally or to cheat on others (e.g., Gneezy, 2005; Cappelen et al., 2017; Gneezy et al., 2018; Kocher et al., 2018; Abeler et al., 2019).

[4] 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.

[5] See, for instance, https://www.netdoktor.de/symptom-checker/ or https://www.secondopinions.com (accessed on 14 December 2022).

[6] In our experiment, we used the following site for Lenovo machines: https://thinkwiki.de/Hauptseite (accessed on 14 December 2022).

[7] See, for example, https://www.jameda.de or https://lawyers.com or, more generally, https://www.yelp.com (accessed on 14 December 2022).

[8] Mimra et al. (2016b) find in a laboratory setting that introducing the possibility to gather costly second opinions significantly reduces the level of overtreatment. This finding is not necessarily in contrast with the finding of Bindra et al. (2021) as consumers in real credence goods markets have always the possibility to gather costly second opinions (i.e., experts are also aware in the BASELINE treatment that consumers could gather a second opinion). In fact, Bindra et al. (2021) addresses the question if revealing that a second opinion has already been gathered reduces the repair price further.

consumer him- or herself.[9] A self-diagnosis retrieved from online sources will almost always remain noisy and it will not transform a consumer into an expert about the credence good. Also, revealing that information has been retrieved from online sources will arguably not convey the information to the expert seller that his recommendation will most likely be accepted. So, the question we are addressing is whether an easily accessible, but noisy self-diagnosis can benefit the consumer *on average*.[10]

Turning to the second channel of information – internet reviews (which are not considered in any paper referenced in the previous paragraph) – 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, 2013, 2018; Bohnet and Huck, 2004). This literature typically investigates sales offers on trading platforms like eBay or Amazon where search or experience goods are offered. With search goods the quality of the good is observed by the customer before the interaction while with experience goods the quality is only learned after consumption or inspection. In contrast to this literature, we are interested in the information content of internet reviews for a credence goods market transaction where consumers are typically not even able to judge *ex post* whether they were provided the good or service that maximized the gains from trade. Although consumers can hardly evaluate the credence attributes involved in a credence goods market transaction, it is typically possible for them to judge dimensions of the transaction that exhibit a search or experience good character (e.g, friendliness or promptness). So, the question is whether those sellers in a credence goods market that are rated higher by consumers – arguably based on search or experience qualities – are also those sellers that defraud consumers less on unobservable credence attributes. To the best of our knowledge, no previous study has addressed this question. Ex ante, it is difficult to judge whether information from rating platforms will help consumers in markets for credence goods. This is not only the case because consumers cannot even ex post judge whether they got the right product or service, but also because reviews on rating platforms may be unreliable as well. In fact, sellers may have incentives to manipulate or fake the ratings themselves or to commission benevolent ones (see Ockenfels and Resnick, 2012).[11] There is abundant literature available that fake reviews promote low quality and that they can lure consumers even into buying such low quality products (Anderson and Simester, 2014; Mayzlin et al., 2014; Luca and Zervas, 2016; Akesson et al., 2022). The manipulation of reviews typically works via inflated star-ratings or exaggerated language in the review texts (Hu et al., 2012). While these channels are well understood in ordinary and experience goods markets, there is no previous literature on the informational value of internet ratings for credence goods markets. Yet, there is some literature addressing the related question of the impact of reputational concerns in the latter type of markets. An early example is the large-scale lab experiment on the influence of institutional and market conditions on the extent of fraud in credence goods markets by Dulleck et al. (2011). In their experimental design the authors have a (private-history) condition, in which they allow consumers to keep track of their own past experience with a particular seller. The authors report that adding this possibility for reputation building increases efficiency thanks to a higher volume of trade, but only when neither liability nor verifiability are in place. In a related lab experiment, Mimra et al. (2016a) find that public information about the past behavior of expert sellers does not necessarily reduce the level of fraud. Turning to experimental evidence from the field, Schneider (2012) investigates in his pioneering study the impact of reputational concerns on the behavior of car mechanics. Based on data from undercover garage visits Schneider (2012) finds hardly any evidence indicating that reputation considerations affect mechanics' provision or charging behavior.[12]

To examine the influence of the two channels through which consumers in credence goods markets can acquire information, we ran two waves of a natural field experiment in computer repair shops in Germany. In both waves, repair shops were presented with a manipulated computer. In the first wave, we varied whether – and, if yes, how – consumers revealed the information retrieved from the internet about the potential source and magnitude of the problem. In the baseline treatment consumers brought the computer to the shop asking for repair without mentioning any supposition about the source of the problem. In one of the alternative treatments mystery shoppers mentioned a vague self-diagnosis of the problem retrieved from an internet page, while in another alternative treatment the shoppers stated a price limit for the repair that corresponded to the self-diagnosis in the first case (without revealing any self-diagnosis, however). It is by no means clear whether one of these ways to reduce the perceived informational advantage of sellers benefits the customer – and if yes, which way benefits the customer most.

By varying how consumers reveal the information retrieved from the internet we can also address whether stating a price limit can be easily interpreted as revealing information about the consumer's willingness to pay.[13] If experts exploit the price limit information for price discrimination, we should see prices close to the stated price limits and we would most likely see different prices depending upon whether consumers state a price limit (which might be interpreted as a willingness to pay) or whether they mention the self-diagnosis (which rather indicates how knowledgeable consumers are than how much they are willing to pay).

[9] There are several other differences between Bindra et al. (2021) and our experiment. First, the treatment variations in Wave 1 of the present paper are implemented via email, while they were implemented by the mystery shoppers in Bindra et al. (2011). This means that the mystery shoppers are blind with respect to the treatment manipulations in the current paper and this avoids having this knowledge bias their perceptions and behaviors, thus impacting the study's internal validity. Second, the opinion of another expert expressed in Bindra et al. (2021) addresses whether the device is in general repairable or not whereas in the present study the self-diagnosis describes the potential source of the problem. Third, the manipulation implemented in Bindra et al. (2021) is harder to diagnose correctly – which is essential as otherwise the second-opinion script would have been unrealistic. The price of this intentional design choice in Bindra et al. (2021) is that incompetence and fraud can no longer be disentangled cleanly (which potentially has a negative effect on internal validity). As we will see later, competence is not really an issue in the current experiments where the rate of successful repairs is 98% (while it is only 75% in Bindra et al., 2021). Finally, in Bindra et al. (2021) a successful repair requires spare parts while with our manipulation no spare parts are needed for the repair. This means that cost differences for the spare parts could drive price differences in Bindra et al. (2021) while this is not the case in the present study, thus removing a potential confound across treatments.

[10] Consumers might have different motives for uttering a specific self-diagnosis. For instance, they could hope that the self-diagnosis helps the expert to find and solve the problem more quickly and this will lead to a reduced bill. Alternatively, they could intend to signal competence so that the expert might be less likely to cheat on them.

[11] 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 14 December 2022).

[12] In addition to the studies that attempt to answer the question of whether the prospect of future interaction can have a fraud-reducing effect on credence goods markets, there is also a literature considering the possibility that reputational concerns might lead to *less* consumer-friendly behavior (see Ely and Välimäki, 2003 and Ely et al., 2008 for formal models). Grosskopf and Sarin (2010) explicitly test for the presence of 'good' as well as 'bad' reputation in a series of lab experiments and find that the positive effects of reputation on market efficiency are generally not as strong as predicted by theory, while the negative effects are basically absent.

[13] The potential influence of revealing information about a consumer's willingness to pay has previously been investigated by Balafoutas et al. (2013) and by Gottschalk et al. (2020). The former study examines this issue in a market for taxi rides by varying consumers' clothes (business outfit versus casual outfit), while the latter addresses it in the market for dental care by modifying clothing and accessories. However, using different clothes as a proxy for one's willingness to pay is a rather vague signal. Our present study is the first one that employs an explicit price signal in a credence goods context.

We find that uttering a vague, but correct, self-diagnosis about the computer's problem does not reduce the average repair price substantially compared to a control condition where the mystery shopper did not mention anything when bringing the computer to the repair shop. However, mentioning a vague, but incorrect, self-diagnosis increases the average repair price substantially. Since the signal generated by a diagnosis software on the internet is almost always noisy and since consumers cannot distinguish between a correct and an incorrect diagnosis, an implication of this result is that revealing a noisy and vague self-diagnosis is a costly mistake for consumers. Somewhat strikingly, we find the same pattern when mystery shoppers indicate a price limit (that reflects either the correct or the incorrect self-diagnosis). Giving a low price limit (in line with the correct self-diagnosis) does not reduce the average repair price substantially compared to the control condition. However, stating a high price limit (in line with the false self-diagnosis) more than doubles the repair price (yet the average repair price still stays far below the stated price limit, which speaks against plain price discrimination of experts). These results imply that uttering a price limit is also, on average, a costly mistake for consumers. We think that these results could be relevant for several other repair markets (e.g., car repairs, mobile phone repairs, household appliances repairs, ...) – provided that it is a problem that is relatively easy for an expert provider to diagnose correctly.

In the second wave of our experiment, we examined the informational value of internet ratings of computer repair shops. Recall that in credence goods markets consumers cannot judge even *ex post* whether they got overtreated or overcharged by sellers, for which reason internet ratings may not have any predictive value for honest service. We test the hypothesis that shops with worse ratings are more dishonest in the sense of charging higher prices for the same service than better rated shops. We find that this is, indeed, the case. A closer look at the content of the ratings reveals that comments about (un)friendliness and (un)successful repairs play the most important role when it comes to determine ratings. Comments about the price play only a subordinate role. This suggests that the observed experience attributes evaluated in consumer reviews are correlated with unobservable credence attributes which finally determine the repair price. Importantly, we can dig even deeper into the predictive power of internet ratings by exploiting Yelp's classification into recommended and non-recommended ratings. The latter are considered by the platform as less reliable. Our data confirms for recommended ratings that 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 are associated with *higher* (rather than lower) prices. Taken together these findings suggest that internet ratings may be a cursed blessing, since they reveal valuable information, but only when they are reliable and unlikely to be manipulated.

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. Before the start of the wave, we bought 12 identical, completely refurbished and perfectly working laptops. In each of the computers, we removed the random-access memory (RAM) modules slightly from their slots (see Online Appendix B for the specification of the laptops and the manipulation in detail). Loose RAM modules are not an exotic problem, they are often the consequence if a laptop drops on the floor. With a loose RAM module, the computer cannot be booted, but produces a black screen and a distinct acoustic error message. Several webpages allow inferring the likely problems from the acoustic messages.[14] In our case, the page suggests a problem with the RAM modules as the first potential cause, and an issue with the main board as the second potential cause. Given that the page suggests two potential causes for not being able to boot the computer, a self-diagnosis solely based on this information is noisy. At the same time the self-diagnosis is also vague because it points only out that there could be problems with specific components without going further into details (e.g, a problem with the RAM modules could imply that one or two modules must be replaced or, as in our case, that they are not properly installed). A consequence of the noisy and vague signal is that for a consumer it remains ambiguous what the real problem is, implying that appropriate service and pricing are hard to identify. By contrast, a repair shop should straightaway be able to diagnose and solve the problem correctly because the loose RAM modules catch the eye of the expert immediately when the computer is opened and because the correct repair for this problem is simply to put the RAM modules back into the slots.

The fact that the problem can easily be diagnosed and repaired by a computer repair shop is an important feature of our experiment because our primary research interest is in intentional fraud, but not in incompetence. 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 (or overcharging in case the additional repair or service is billed but not provided). Second, the costs for a proper repair only include working time, since no spare parts are necessary, and are thus rather low. The low costs imply 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 then randomly assigned (with the help of a random number generator) a treatment and an undercover helper to each selected shop, subject to having 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. Initially, we wrote an email (from a private address) to a repair shop with the following text (originally in German, as was all communication):"*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 had confirmed that we could bring the laptop we sent the actual treatment variation in a second email. Afterwards we sent one of our mystery shoppers with the computer to the shop. Importantly, the mystery shoppers were unaware of our research question, the treatment variations and the treatment to which a specific shop had been assigned to. Moreover, our mystery shoppers were instructed simply to drop off the computer at the repair shop and keep the interaction as short as possible (to minimize any confounds from personal communication). The following five treatments were implemented:

- BASELINE: Here, 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:*

[14] See, for example, https://thinkwiki.de/Error_Codes (accessed on 14 December 2022).

"veronika123". Please inform me as soon as you know more." We neither mentioned any potential source of the problem, nor any limit for the repair price.

- CORR-GUESS: 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 with the RAM modules. Maybe this helps.*" In this case the uttered conjecture about the problem is vague but 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 CORR-GUESS than in BASELINE. However, given that the conjecture is vague, it still leaves room for a dishonest expert seller to overtreat or overcharge the customer (e.g, by replacing RAM modules or claiming a replacement without doing so). Therefore, it remains an empirical question whether prices in CORR-GUESS are lower than in BASELINE.[15]
- CORR-PRICE: This treatment also started with the script from BASELINE, but then the following text was added: "*If the cost of the repair is below 50 Euro, please do the repair. If the cost of the repair will exceed 50 Euro, please contact me again.*" This price limit corresponds approximately to the estimated costs if the vague but correct conjecture from CORR-GUESS is the true problem (which is actually the case).[16] Please note, however, that in CORR-PRICE we did not mention any kind of self-diagnosis.
- INCORR-GUESS: Again, 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 with the main board. Maybe this helps.*" This conjecture is vague and false, but recall that the acoustic error message suggested a problem with the main board as a potential source of the problem. This means that our script conveys a realistic scenario, and does not necessarily reveal incompetence on the side of the consumer (which shops might be tempted to exploit). It is important to note that this treatment variation was deliberately designed in a way that the vague and false conjecture should be easily detected by every repairer. So, even if a repair shop took the false conjecture as the starting point, generating a correct diagnosis and repairing the laptop should not lead to a meaningful increase in working time, meaning that the price an honest shop charges in INCORR-GUESS should not be higher than the price charged in BASELINE. However, the false conjecture arguably generates more room for a dishonest expert seller to overtreat or overcharge the customer (e.g., experts could anticipate less negative consumer reactions in such cases). Assuming that some shops exploit this opportunity, we expected higher prices, on average, in INCORR-GUESS than both in BASELINE and in CORR-GUESS.
- INCORR-PRICE: In this treatment we added the following text after the initial script from BASELINE: "*If the cost of the repair is below 200 Euro, please do the repair. If the cost of the repair will exceed 200 Euro, please contact me again.*" The motivation for this price limit is that it corresponds approximately to the estimated repair costs if the vague and false conjecture from INCORR-GUESS would be the true problem. However, in the present treatment only the price limit is communicated, but not a combination of the price limit and the false conjecture.

[15] We want to emphasize that a correct guess without the element of vagueness in combination with the triviality of the manipulation would be unrealistic because in this case the consumer could simply put the RAM modules back in the slots himself and a shop visit would be superfluous.

[16] If a shop called us and mentioned an estimated price exceeding the limit of 50 Euro, we always asked the shop to proceed with the repair. The same applies to treatment INCORR-PRICE below.

The first wave of our experiment was conducted between November 2015 and July 2016. We implemented all treatments of this wave simultaneously in order to exclude the possibility that seasonal effects drive the treatment differences. We sent seven undercover helpers ("mystery shoppers") during regular opening hours to the repair shops on our list. As indicated above, treatment assignment was random and mystery shoppers were blind to the treatment. Our randomization strategy and selection of mystery shoppers was successful as we observe no significant differences in the observable characteristics of the repair shops across treatments (see Table A1 in the Appendix) and also no significant effects of individual mystery shoppers on the repair price (see below).

To receive additional information on how realistic the treatment scripts are, on the channels through which an (in)correct guess might affect the repair price and on the question whether the implemented manipulation is really that easy to diagnose, we conducted a survey in November 2022 in 30 repair shops (about one-third of the shops were also part of Wave 1 of our main natural field experiment).[17] In this survey we elicited the commonness of stating a self-diagnosis (Q1); the difficulty to diagnose the issue with the laptop correctly (Q2); whether an incorrect guess about the source of the problem is likely to increase the repair time (Q3); whether a correct guess about the source of the problem is likely to decrease the repair time (Q4); and the commonness of stating a price-limit (Q5).[18] The interviewees were asked to answer these questions on a scale from 1 to 5 (see the end of Appendix B for the exact wording of the questions and the corresponding answer categories). Table 1 presents the means and standard deviations for the answers to each question. Regarding realism of scripts, in Q1 17 out of 30 shops indicated that customers state very often or often a rough self-diagnosis and in Q5 11 out of 30 shops indicated that customers state very often or often a price limit when asking for a repair. When it comes to Q2 regarding the difficulty of diagnosing the problem correctly, 28 out of 30 shops answered that the problem is very easy or easy to diagnose. Similarly, for Q3, 26 out of 30 shops stated that an incorrect guess does certainly not or rather not lead to an increase in diagnosis and repair time. However, in Q4 a smaller proportion of shops (19 out of 30) answered that a correct guess does certainly not or rather not lead to a decrease in diagnosis and repair time. In sum, the survey suggests that the treatment manipulations reflect fairly realistic and typical situations and the substance of our communication with the shops is used by a sizeable fraction of customers out in the field. More importantly, the survey provides further evidence that the manipulated laptops are really easy to diagnose correctly and that there is no reason to assume that an (in)correct guess leads to a justified (increase) decrease in diagnosis and repair time.

3. Results of Wave 1 – The effects of a noisy self-diagnosis and of price limits

In total, we collected 119 observations; 24 for BASELINE, 24 for CORR-GUESS, 24 for CORR-PRICE, 23 for INCORR-GUESS and 24 for INCORR-PRICE. Out of these shops, one shop in INCORR-GUESS and one shop in CORR-PRICE claimed that the computer could not be repaired.[19] We exclude these shops from the analysis, since no service was provided in those cases and consequently, there is no 'repair price' which could be compared in a meaningful way. This leaves us with a total of 117 shops that provided service; 116 of

[17] We thank an anonymous referee for suggesting this survey.

[18] In (Q2) we explained the issue with the laptop but did not say that we had manipulated the device.

[19] The shop in INCORR-GUESS charged 20 Euro for the diagnosis and the shop in CORR-PRICE suggested to buy a new computer in the shop and charged 29 Euro for the diagnosis.

Table 1
Survey regarding the commonness/effect of the (treatment) manipulations.

Question	Q1: Commonness of stating a self-diagnosis (1 very often – 5 very rare)	Q2: Difficulty to diagnose the manipulated laptops (1 very easy – 5 very hard)	Q3: Increased repair time after incorrect guess (1 certainly not – 5 for sure)	Q4: Decreased repair time after correct guess guess (1 certainly not – 5 for sure)	Q5: Commonness of stating a price-limit (1 very often – 5 very rare)
Mean	2.43	1.53	1.63	2.43	3.47
Standard deviation	1.21	0.63	1.07	1.65	1.48

them were able to repair the computer, and one shop claimed that the computer is irreparable and saved the data on an external hard drive.[20] Given that practically all computers were successfully repaired, the repair price is our main variable of interest.

On average, the repair price is lowest in BASELINE (38.21 Euro), and only slightly higher in the treatments with a vague but correct guess or a low price limit (44.85 Euro in CORR-GUESS and 47.09 Euro in CORR-PRICE), but it practically doubles when the computer owner utters a vague and wrong conjecture or a high price limit (84.50 Euro in INCORR-GUESS and 86.86 Euro in INCORR-PRICE). The repair price in INCORR-GUESS is affected by a massive outlier of 450 Euro and the average price reduces to 67.10 Euro without this outlier. Given the magnitude of this outlier, we exclude it in the subsequent figures and regression models and present the same regressions including the outlier in Appendix A.[21] Fig. 1 shows the distribution of the repair prices with the help of box plots for each treatment, showing that the variance of repair prices is particularly larger in INCORR-GUESS and in INCORR-PRICE than in the other three treatments. In addition, Fig. 2 presents the cumulative distribution functions for each treatment.[22] It is interesting to mention that in CORR-PRICE we observe about one third of prices above the stated price limit, indicating that stating a price limit need not protect consumers from relatively high prices.

Column [1] of Table 2 presents an OLS-regression with the repair price as the dependent variable and dummies for the treatments CORR-GUESS, CORR-PRICE, INCORR-GUESS and INCORR-PRICE. Column [2] of Table 2 adds in addition the following controls as independent variables: "One-man business" controls for whether a shop is run by a single person or not (this information was taken from the homepage of the respective shop and we double checked this information with the help of our mystery shoppers when they visited the shops). If a shop is run by a single person, the owner is arguably the 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. "Number of competitors" measures the number of repair shops within a circle of 5 km. This variable accounts for potential competition effects. Finally, the variable "Rental price" controls for the prices of apartments in the district where a repair shop is located.[23] 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 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.

In Table 2, we take BASELINE as the benchmark and find that CORR-GUESS and CORR-PRICE increase prices, but not significantly so. The latter two treatments are statistically indistinguishable from BASELINE (according to a Kolmogorov-Smirnov-test). From this we can infer that uttering a vague but correct self-diagnosis about the problem or a price limit in line with that self-diagnosis does *not reduce* repair prices compared to BASELINE.

At first sight, one might conjecture that the null-effect of the two CORR interventions is due to hardly any mistreatment in BASELINE, so that there might be almost no room for improvement in CORR-GUESS and CORR-PRICE. In order to assess this conjecture, we first have to discuss what prices could be considered as justified and what prices could be indicative of mistreatment. As most shops in the German computer repair market charge a diagnosis fee (which is most of the times offset in case of a repair order), this diagnosis fee could be considered as a natural candidate for a justified repair price: Once the problem is diagnosed correctly (which is actually fairly easy and should happen straightaway), repairing it (i.e., putting the RAM modules back in the slot) requires no additional time and effort and no spare parts. Therefore, charging a price in addition to the diagnosis fee seems to be unjustified and is indicative of mistreatment. For a majority of shops (N = 62) we could identify the diagnosis fee which was on average 36.47 Euro. In BASELINE, 42 % of the shops charge a price above the average diagnosis fee and therefore, we conclude that roughly 40 % of the shops charge prices that are indicative of mistreatment. In addition, the observed variance of repair prices in BASELINE is noticeable and not condensed around the average price of 38 Euro, with a minimum of 0 Euro, a maximum of 96.25 Euro, and with a quarter of shops charging 15 Euros or less and another quarter charging more than 59.50 Euro.

Altogether, these figures suggest that the credence goods problem in BASELINE is not pervasive but still existing and there would be definitely room for improvement.[24,25] Therefore, the self-diagnosis in CORR-GUESS is probably too vague (e.g., a shop could still claim that one RAM module is broken and needs to be replaced). At the same time, the price limit in CORR-PRICE is probably too high (i.e., it is above the average repair price in BASELINE), but also not binding (since about one third of shops charge prices above the price

[20] We include the latter shop since service was provided (i.e., the shop billed a new hard drive and the working time for saving data on this hard drive). Results remain unchanged if we dropped this shop.

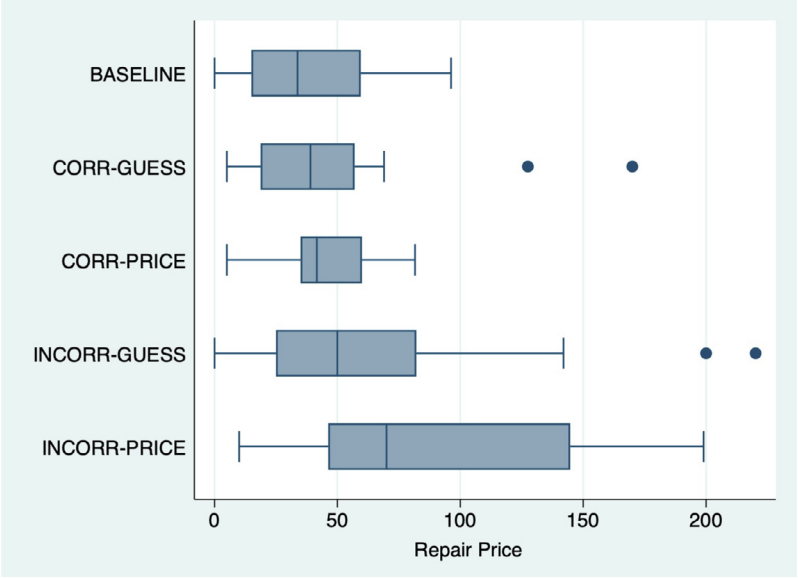
[21] The main difference between the models with and without the outlier is that in the specifications without the outlier the coefficient for the number of positive ratings is significant and this is not the case when the outlier is included in the analysis (see Tables A7 and A8).

[22] In addition, Table A2 in the Appendix shows the min, max, mean, 10th, 25th, 50th, 75th, and 90th percentile for each treatment.

[23] This information was taken from https://www.wohnungsboerse.net. We had to use rental prices for apartments as a proxy because rental prices for business premises were not readily available.

[24] That the null effect is not due to the absence of fraud in our BASELINE condition is also suggested by the findings in Hall et al. (2019). They measure the level of fraud in a credence goods environment very similar to the one in our BASELINE treatment. This is done by comparing the average repair price across two environments: In the credence goods environment the mystery shopper enters the repair store with a broken device and asks for a repair, mentioning that he has no idea which kind of repair is needed to fix the problem (like in our BASELINE treatment). In the ordinary goods environment the mystery shopper enters the store with a broken device and asks to change the specific part that causes the problem. The authors find that prices are about 40% higher in the credence goods environment.

[25] These figures are also in line with the results of various laboratory experiments on dishonesty suggesting that most subjects are partial liars (see, for example, Fischbacher and Föllmi-Heusi, 2013 and Mazar et al., 2008).



Each box plot displays the minimum (beginning of the left whisker), first quartile (left whisker), median (vertical line in the box), third quartile (ends at the right-hand side of the box) and maximum (end of the right whisker). In addition, outliers (i.e., observations greater than 1.5 times the interquartile range which is defined as the distance between Q3 and Q1) are indicated by dots.

Fig. 1. Box plots of the repair price for each treatment.

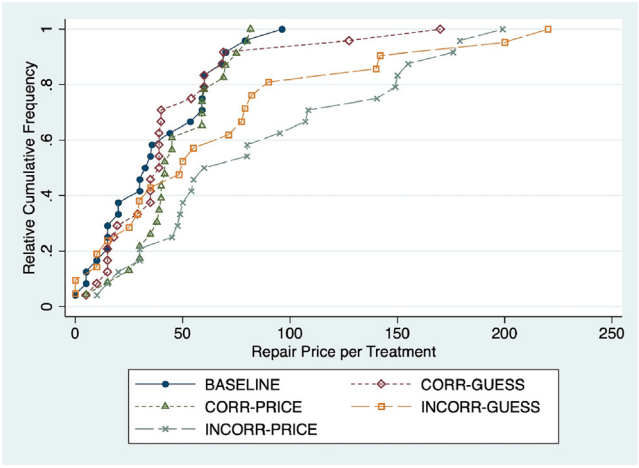


Fig. 2. Cumulative distribution function of repair prices, conditional on treatment.

limit) in order to reduce the repair price below the BASELINE benchmark.[26]

Contrary to the null-effect of a *vague but correct* conjecture or a *correct* price limit, however, the treatment effects of INCORR-GUESS and INCORR-PRICE are economically large and statistically highly significant. The estimated price difference between BASELINE and the INCORR-GUESS treatment is about 29 Euro and the estimated price difference between BASELINE and the INCORR-PRICE treatment is about 49 Euros. Given an average repair price of 38 Euro in BASELINE, this implies an estimated price increase of about 76 % (129 %) in case the consumer utters a vague and false self-diagnosis about the problem (states a too high limit for the repair price). These results are even more pronounced when taking into account fixed effects for our mystery shoppers (see column 1 of Table A3 in the Appendix) and when including our control variables that turn out to be insignificant (see column 2 of Table 2). Looking at what shops claimed to have repaired (other than putting the RAM-modules back into their slots), helps to understand where the differences between the treatments come from. Table A4 in the Appendix lists the 15 cases where shops claimed services or repairs that are not related to our manipulation. First, it is noticeable that such unnecessary claims were made only in 5 out of 71 cases (7 %) in the set of treatments BASELINE, CORR-GUESS and CORR-PRICE, while this happens in 10 out of 46 cases (22 %) in treatments INCORR-GUESS and INCORR-PRICE. So, additional (and unnecessary) services become significantly more likely in the latter two treatments ($p < 0.05$; χ^2-test). Second, the excess repairs are significantly cheaper in the 5 cases of the first set than in the 10 cases of the second set of treatments (74 Euro vs 174 Euro on average; $p < 0.01$; Mann-Whitney U test). We summarize our first results and their implications as follows:

Result 1 and Implication 1: *Uttering a correct but vague conjecture about the problem or stating a low price limit that is in line with the correct conjecture does not reduce the repair price on average. However, a vague and incorrect conjecture or a too high price limit increases the average price substantially. Since the diagnosis on the basis of information retrieved from the internet is almost always noisy – and since consumers cannot distinguish between a correct and an incorrect diagnosis – an implication of Result 1 is that mentioning a*

[26] We have to admit that our experiment is underpowered to detect small price reducing effects of the CORR-GUESS and the CORR-PRICE treatment. However, our data reveals absolutely no tendency in the direction of such a price reducing effect. One may wonder whether price limits that are clearly below the expected repair price might help the consumer. In our design, such a treatment was difficult to implement, given that the average BASELINE repair price was already fairly low. Yet, we have collected some additional data on the effects of a very low price limit using desktop computers and a manipulation that caused an average repair price of 189 Euro in the baseline treatment (see Bindra et al., 2021, for a detailed description of the manipulation and the procedures in the baseline treatment). When we then implemented a treatment where a mystery shopper stated a price limit of 100 Euro – above which the repair shop should call before actually doing the repair (in which case the mystery shopper always asked for the repair) – we found an average repair price of 221 Euro, which was far above the stated price limit and even higher than the average baseline price. This result suggests that even a price limit that is significantly below the repair price of a corresponding baseline treatment has no positive effect for the consumers.

Table 2
Regression analysis of repair prices without outlier.

Dependent variable (OLS regressions) Independent variables	*[1]* *Repair price* *(in Euro)*	*[2]* *Repair price* *(in Euro)*	*[3]* *Repair Price* *(in Euro)*
CORR-GUESS treatment (1 = yes)	6.64	8.74	13.66
	(9.30)	(10.43)	(10.73)
CORR-PRICE treatment (1 = yes)INCORR-GUESS treatment	8.87	12.05	15.44*
(1 = yes)	(6.83)	(7.58)	(7.90)
	28.89**	32.24**	37.87**
	(14.60)	(15.66)	(15.71)
INCORR-PRICE treatment (1 = yes)	48.65***	54.52***	59.06***
	(12.94)	(13.66)	(12.53)
One-man business (1 = yes)		14.38	17.45*
		(9.50)	(10.10)
Number of competitors within 5 km		0.01	0.11
		(0.62)	(0.63)
Rental price in the district of the shop (€/m^2)		0.93	0.31
		(0.95)	(0.96)
Constant	38.21	18.67	19.62
	(5.38)	(14.66)	(14.85)
Negative ratings (log of number of ratings with 1 star or 2 stars plus 1)			22.79***
			(6.38)
Positive ratings (log of number of ratings with 3 stars or better plus 1)			−6.52**
			(2.60)
# Observations	116	116	116
R-squared	0.15	0.18	0.25

OLS-regressions (robust standard errors) with repair price (in Euro) as dependent variable, including, as explanatory variables, a dummy for CORR-GUESS, a dummy for CORR-PRICE, a dummy for INCORR-GUESS, a dummy for INCORR-PRICE, a dummy for being a one-man business, the number of other shops within a radius of 5 km, an index for the average rental price in the district of the shop and the log of number of negative/positive ratings. ***, **, * denote significance at the 1 %, 5 %, 10 % level, standard errors in parentheses.

noisy self-diagnosis or the corresponding price limit is, in expectation, a costly mistake for in our setting with a simple problem.

Only after having concluded wave 1 of our field experiment, we realized that we could look *ex post* into the internet ratings of the repair shops that we had consulted. 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. So, we collected the internet ratings of our 117 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 A5) that the results do not change qualitatively if we would classify all ratings with 1, 2, or 3 stars as negative, and those with 4 or 5 stars as positive.[27]

Column [3] of Table 2 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 independent variables already included in column [2]. The linear-log model (in which the dependent variable – the repair price – is not transformed, but the independent variables – the number of positive and negative ratings – are logarithmized) is chosen because the effect of ratings on the repair price is expected to retain the same sign (positive or negative) independently of the number of ratings, but the size of the impact is expected to be decreasing in the number of ratings. Such models are often applied when analyzing the effects of reputation on eBay (see, for example, Melnik and Alm, 2002).

We see that negative ratings are associated with significantly higher repair prices and positive ratings are associated with significantly lower repair prices. Adding the ratings changes a few of the other results from column [2]. First, the dummy for CORR-PRICE turns weakly significantly positive, suggesting that stating a price limit that is in line with the correct conjecture is a costly mistake not only in expectation, but even for the (best) case where the conjecture is correct. Second, the dummy for a one-man business becomes weakly significantly positive in column [3], supporting the hypothesis that shop owners who are the full residual claimants charge higher prices.[28]

An interesting aspect not addressed in Table 2 is the question whether the poorly rated shops are those who exploit their customers most in the INCORR-GUESS and the INCORR-PRICE treatment. This is, in fact, the case, as we show in Table A6 in the Appendix. There we find a significantly positive and large interaction effect of bad ratings and a category INCORRECT that pools treatments INCORR-GUESS and INCORR-PRICE. This result suggests that shops with negative ratings are indeed shops that abuse their (perceived) informational advantage – and not simply shops that are expensive (for whatever reason).

4. An ex-ante test of the informational value of internet ratings in credence goods markets – 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

[27] In wave 2 of the field experiment we will introduce separate variables for the Yelp and Google reviews and further differentiate between recommended and not recommended ratings when it comes to the Yelp ratings. For wave 1 this is infeasible, as we have not enough observations to include these subcategories.

[28] The dummy for one-man businesses is also significant in our specification with shopper fixed effects (Table A3 in the Appendix).

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 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.[29] 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 Berlin, 58 shops had 3 or more reviews. We sent our mystery shoppers to each of these 58 shops with a request for a repair.[30] Unknown to the mystery shoppers, 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 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

Relationship between Internet Ratings and Repair Prices We first look at the repair prices of shops, contingent on their average rating being above or below the median rating. 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). Fig. 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, we consider it as striking 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 in our setting: Shops with an average rating worse than the median charge on average significantly higher prices than shops with above median ratings. 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 after consumption.*

Result 2 supports the idea that consumer reviews can provide valuable information even in markets with informational asymmetries that persist after using the service or good. This raises questions about the underlying mechanism because, from a theoretical perspective, consumers do not have the ability to evaluate the quality (and necessity) of a credence goods service. Therefore, it seems plausible to assume that consumers are likely to judge the more easily accessible experience aspects of the transaction, like friendliness or promptness.[31]

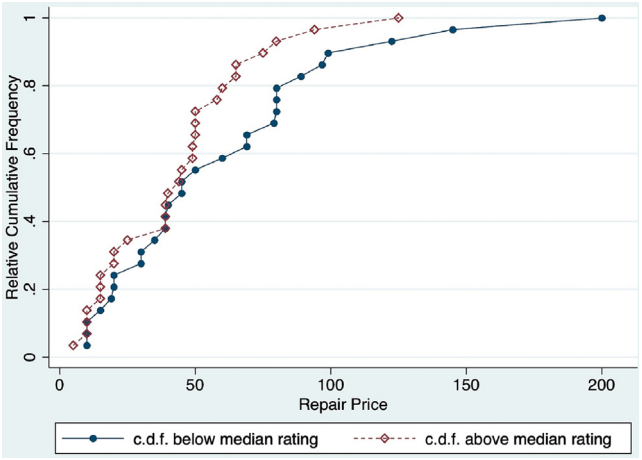


Fig. 3. Cumulative distribution function of repair prices, conditional on average rating above or below median.

Predictive Power of Verbal Comments in the Ratings for the Star Ratings To examine the conjecture that consumers are likely to base their star ratings on the experience aspects of the transaction, we conducted a content analysis of consumer reviews of our shops in wave 2. For the content analysis we hired two research assistants that were otherwise not involved in this project. In a first step, the research assistants analyzed the verbal comments of a random sample of the online reviews for the shops in our data base to get an idea which issues were addressed most often. Following this, the research assistants agreed to code the consumer reviews falling into the following categories: price (cheap or expensive), friendliness (friendly or unfriendly), competence (high or low), success (successful repair or unsuccessful repair) and promptness. The coding was then done independently by the two research assistants and both coded every consumer review of our sample he/she could find. In total, the first research assistant coded 3,077 reviews and the second assistant coded 2,932 reviews. Based on the analysis of both research assistants, 71 % of the reviews exhibit a verbal comment from at least one of the above categories. Further, positive attributes (like cheap, friendly, high competence and successful repair) are mentioned by far more frequently in the reviews than negative attributes (like expensive, unfriendly, low competence and unsuccessful repair). The respective fractions are 26 % for cheap (4 % for expensive), 36 % for friendly (5 % for unfriendly), 33 % for competent (3 % for incompetent), 25 % for success (5 % for no success) and 22 % for promptness. As the shops in our sample are rated quite well on average (see Table 5 below), it is not surprising that the corresponding ratings exhibit more positive attributes than negative ones. In order to assess the predictive power of the verbal contents for the ratings, we focus on the subsample of ratings with verbal contents and regress the ratings (number of stars from $1-5$) on dummies for the attributes from the above categories. The idea of this regression is to check whether negative (positive) attributes are really associated with negative (positive) coefficients as excepted. Table 3 shows the results for the two research assistants separately: all the coeffi-

[29] 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 (by including only shops with more ratings we would have lost too many observations for a meaningful power analysis), 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.

[30] In wave 2, we employed five mystery shoppers. The regression in Table A9 of Appendix A controls for mystery shopper fixed effects and shows that the characteristics of the individual mystery shopper have no significant effect on the repair price.

[31] For example, consider one of our shops that diagnosed a hairline crack in the mainboard and charged 199 Euro for infrared soldering this crack. A naïve customer is most likely not able to evaluate the credence attributes of this transaction – that is, whether the problem was diagnosed correctly and which repair was actually provided. Still, the customer can credibly evaluate the experience attributes of the transaction like how she was treated and how quick the repair was done.

Table 3
Content of reviews and star ratings.

	Research assistant 1	*Research assistant 2*
Dependent variable (OLS-regressions)	*Star rating*	*Star rating*
Independent variables		
	(from 1 – 5)	*(from 1 – 5)*
Cheap (1 = yes)	0.20***	0.27***
	(0.03)	(0.07)
Expensive (1 = yes)	−0.91***	−2.08***
	(0.15)	(0.23)
Friendly (1 = yes)	0.34***	0.21***
	(0.07)	(0.06)
Unfriendly (1 = yes)	−1.57***	−2.56***
	(0.14)	(0.22)
Competent (1 = yes)	0.30***	0.23***
	(0.05)	(0.06)
Incompetent (1 = yes)	−0.71***	−1.09***
	(0.08)	(0.22)
Successful repair (1 = yes)	0.44***	0.27***
	(0.05)	(0.08)
Unsuccessful repair (1 = yes)	−1.45***	−2.89***
	(0.15)	(0.18)
Comments on promptness (1 = yes)	−0.16***	0.12***
	(0.06)	(0.03)
Constant	4.05***	4.43***
	(0.11)	(0.11)
# Observations (reviews)	2,492	1,801
R-squared	0.77	0.68

OLS-regressions (standard errors clustered at the shop level) with star rating (ranging from 1 – 5) as dependent variable, including, as explanatory variables, dummies for the attributes cheap, expensive, friendly unfriendly, competent, incompetent, successful, unsuccessful and promptness.
***, **,* denote significance at the 1%, 5% and 10% level, standard errors in parentheses.

cients are highly significant and point in the expected direction as the dummies for expensive, unfriendly, incompetent and no successful repair decrease the star rating and the dummies for cheap, friendly, competent and successful repair increase the star rating. The coefficient for promptness is negative in one regression and positive in the other, suggesting that comments in this category are not clearly associated with a negative or positive connotation. Given that the coefficients for all dummies in Table 3 are highly significant, we conduct – in Table 4 – a dominance analysis in order to assess the relative importance of the various attributes on the star ratings (see Grömping, 2007, for a discussion).[32] The corresponding general dominance statistics (one for each attribute dummy and each research assistant) are an additive decomposition of the R-squared associated with the full model of Table 3 and can be compared to one another. Column 3 of Table 4 shows the results for the data of the first research assistant: the dummy for "unfriendly" exhibits the highest dominance statistic overall followed by the dummy for "unsuccessful repair". The dummy for "friendly" has the highest dominance statistic amongst the positive attributes followed by the dummy for "competent".[33] Column 5 of Table 4 shows the results for the second research assistant: the dummy for "unsuccessful repair" exhibits the highest dominance statistic overall followed by the dummy for "unfriendly". The dummy for "cheap" has the highest dominance statistic amongst the positive attributes followed by the dummy for "friendly". So, the combined data of the research assistants suggest that the experience attributes "friendly"/"unfriendly" are amongst the most important attributes when it comes to determine good and bad ratings in our sample. Further, it seems that the experience attribute "unsuccessful repair" plays also an important role overall whereas comments about the competence of the expert and the price of the repair do so only within the domain of positive attributes.

This supports our conjecture that the online ratings of credence goods sellers are predominantly based on experience characteristics and we therefore interpret our Result 2 as evidence indicating that those sellers that are rated higher based on experience characteristics are also those sellers who defraud consumers less on the credence goods dimension of the transaction.[34]

The Adverse Effects of Manipulated Ratings In this subsection, we dig even deeper into the informational value of consumer reviews. After all, sellers have incentives to get good reviews as they 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. However, 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 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 in our experiment. In fact, 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 repair 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."[35] This means that Yelp tries to filter out fake reviews or reviews from reviewers who have a poor reputation in the community.

Table 5 presents the average number of recommended and non-recommended reviews for the 58 shops and the distribution of ratings from 1 star to 5 stars.[36] Column [1] shows data for recommended reviews, and column [2] for non-recommended reviews on Yelp. The first thing to notice is that only 23 % of the total number of reviews are recommended by Yelp. A second important observa-

[32] Dominance analysis is an ensemble method in which importance determinations about independent variables are made by aggregating fit metrics across multiple models where each combination of independent variables is considered. We present in Table 4 general dominance statistics based on the R-squared. General dominance statistics have the advantage that they distill the entire ensemble of models estimated into a single value for each independent variable that can be compared to one another to determine the relative importance of the variables. This means that the dominance statistics reported in Table 4 are based on 511 different regression models and therefore, it is not necessarily the case that the relative magnitude of the coefficients reported in Table 3 are in line with the rank of the general dominance statistics reported in Table 4.

[33] For example, the value of 0.185 for the dummy "unfriendly" means that, on average, this attribute results in an increment to the R-square of 18.5% when it is included in the model shown in Table 3. This attribute is therefore relatively more important than the attribute "friendly" where the dummy has only a value of 0.073.

[34] From a personality psychology perspective this does not sound unreasonable. For example, the honesty-humility dimension of the HEXACO model of personality structure is not only a valid predictor for cheating behavior in various field settings, it is also related to other dimensions (like agreeableness) and thus to one's propensity to cooperate or to hold one's temper (and thus be friendlier than when losing one's temper). See, e.g., Ashton and Kibeom (2005) or Hillbig and Zettler (2015).

[35] See, for example, the "Not Recommended" section on the following webpage: https://www.yelp.com/biz/notebookservice030-berlin-3?osq=laptop+reparatur (accessed on 14 December 2022).

[36] These data were retrieved from the rating platforms in January 2018.

Table 4
General dominance statistics.

	Research assistant 1		*Research assistant 2*	
Dependent variable (Star rating) *Independent variables*	*Dominance Statistic*	*Ranking*	*Dominance Statistic*	*Ranking*
Cheap	0.028	8	0.038	5
Expensive	0.075	4	0.127	3
Friendly	0.073	5	0.031	6
Unfriendly	0.185	1	0.192	2
Competent	0.070	6	0.025	7
Incompetent	0.100	3	0.042	4
Successful repair	0.060	7	0.019	8
Unsuccessful repair	0.175	2	0.193	1
Comments on promptness	0.003	9	0.014	9
# Observations (reviews)	2,492		1,801	
R-squared	0.767		0.680	

General dominance statistics (based on 511 regressions) associated with the full model from Table 3.

Table 5
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)		
#⭐	0.57 (0.57)	0	5	2.02 (5.05)	0	33	3.66 (4.87)	0	27
#⭐⭐	0.12 (0.42)	0	2	0.26 (0.95)	0	5	0.71 (1.33)	0	7
#⭐⭐⭐	0.28 (0.81)	0	5	0.24 (0.66)	0	4	0.59 (1.33)	0	8
#⭐⭐⭐⭐	0.74 (1.46)	0	6	0.90 (2.61)	0	17	1.72 (4.50)	0	32
#⭐⭐⭐⭐⭐	1.98 (2.93)	0	15	10.79 (23.72)	0	109	19.5 (24.40)	0	153

Standard deviation in parentheses.

tion is that the average rating of recommended reviews (3.94) is lower than the average rating of non-recommended reviews (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 (with 76 %) than in the recommended reviews (with 54 % of 5-star ratings).[37] In column [3] of Table 5, we show the reviews from Google for the 58 shops. Google does not offer a distinction between recommended and non-recommended shops, however.

In Table 6, 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 2. With regard to the ratings, we distinguish between recommended and non-recommended reviews on Yelp, and keep Google reviews as a separate category.[38]

The first two explanatory variables in Table 6 draw on recommended reviews on Yelp and they reveal a pattern that matches our hypothesis: Negative ratings are associated with significantly higher repair prices, while positive ratings correlate significantly with lower prices. This confirms in the *ex ante* setting of wave 2 the previous result that we had obtained *ex post* from wave 1. Yet, based on the distinction between recommended and non-recommended reviews we are now able to examine in a more refined way which types of 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 are associated with significantly *higher* prices – which is exactly opposite to the effect of recommended positive reviews on Yelp. We consider the latter an important, 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.[39]

Reviews on Google are associated with lower prices when they are positive, but have no significant effect when they are negative.

[37] This pattern matches earlier observations by Hu et al. (2012) very well.

[38] In Tables A10 and A11 in Appendix A we show alternative specifications: In Table A10, ratings with 1, 2, or 3 stars are classified as negative, while those with 4 or 5 stars are classified as positive. In Table A11, we exclude as a further robustness check the Google reviews because of potential multicollinearity issues (with the Yelp reviews, in case a reviewer posts a rating on both platforms). The qualitative results remain unchanged.

[39] 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 (for correlational evidence on this possibility see https://www.nytimes.com/2011/05/22/your-money/22haggler.html?_r=2; link accessed on 14 December 2022). If correctly identified as fake reviews, this kind of unfair competition could lead to a significantly negative correlation between non-recommended negative ratings and repair prices. We do not find such an effect, however.

Table 6
Recommended and non-recommended reviews and repair price.

Dependent variable (OLS-regressions) *Independent variables*	*Repair price (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 5 km	−1.04
	(1.23)
Rental price in the district (€/m^2)	−0.97
	(3.21)
Constant	99.64***
	(36.02)
# Observations (repair shops)	58
R-squared	0.29

OLS-regressions with repair price (in Euro) as dependent variable, including, as explanatory variables, positive and negative ratings of recommended and non-recommended reviews on Yelp, and on Google (where there is no distinction between recommended and non-recommended reviews). Additional controls like in Table 2 apply.
***, **,* denote significance at the 1%, 5% and 10% level, standard errors in parentheses.

The latter result is in contrast to the finding that recommended negative reviews on Yelp are associated with significantly higher prices, but in line with the non-significant impact of non-recommended negative reviews on that platform. So, compared to recommended Yelp reviews, reviews on Google seem to have 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.[40]

The three additional controls at the bottom of Table 6 remain insignificant, which is partly different from the findings in column [2] of Table 2. A potential explanation for the insignificance of these control variables is that in Table 6 we capture more information from the ratings than we were able to do in Table 2 – 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 follows:

Result 3 and Implication 3: *The informational value of internet ratings is heterogeneous in our setting: the most informative ratings are recommended ones (on Yelp) where negative ratings are associated with significantly higher prices and positive ratings are associated with significantly lower prices; for non-recommended positive reviews the correlation is exactly reversed (they are associated with higher prices), while non-recommended negative reviews are not informative for prices. Together these results suggest that consumers in credence goods markets can benefit from internet reviews – but 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, Uber, 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, 2018). Digitization has also affected labor markets by making extremely flexible work contracts with extraordinarily adjustable working hours feasible (like with Uber; see 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, 2010, 2012; Kerschbamer and Sutter, 2017). The question whether and to which degree consumers on credence goods markets can benefit from easily accessible online information is therefore a crucial one, but it has rarely been addressed so far in the literature.

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. Our main research interest was the causal link between information retrieved from the internet and social media and the extent to which consumers are cheated upon by sellers on these 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, on average, a costly mistake when they acquire a noisy self-diagnosis from specialized webpages about the problem of their computer *and* reveal this self-diagnosis or an associated price limit to the repair shop: Uttering a correct conjecture about the potential problem or an appropriate price limit for that conjecture does not reduce the repair price in comparison to a situation where the computer owner simply asks for a repair; however, an incorrect conjecture or a too high price limit increases the average repair price substantially.[41] A corresponding policy implication would be that consumers in credence goods markets should avoid to reveal their superficial knowledge when handing in a good for repair and authorities should avoid to provide means that transfer superficial knowledge to consumers in credence goods markets. Another advice for customers would be to avoid to give price limits for repairs.

In the second wave of our field experiment, however, we have found encouraging news, as internet ratings about repair shops are clearly associated with repair prices. Shops with better internet ratings charge significantly lower prices (for the same and successful repair) than shops with worse ratings. From a theoretical perspective, the observed correlation between ratings and prices could also have been the other way around as shops with better ratings could, in principle, charge a premium on their services. In fact, such a positive correlation is often observed in non-credence goods markets, like in Ebay auctions for identical products (see, e.g., Melnik and Alm, 2002). While the positive correlation found in these studies is consistent with a reputation-milking story where the causality goes from good reputation (based on experience attributes like communication qual-

[40] See the elaborate discussion of Joy Hawkins on https://searchengineland.com/yelp-vs-google-how-do-they-deal-with-fake-reviews-307332 (accessed on 14 December 2022).

[41] Of course, it is feasible that uttering a correct conjecture has a beneficial effect on the repair price in a context where the problem is more complex and harder to diagnose correctly.

ity, delivery time, accurate description of the good, ...) to higher prices, the negative correlation between ratings and prices found in the present study is consistent with a reputation-building story where honest service leads to a better reputation. What could be the reason that previous studies have presented evidence in line with reputation milking in ordinary and experience goods markets, but that we find evidence in line with reputation building in the present market for credence goods (although the potential for reputation milking in credence goods markets would be even greater)? One potential explanation is that the value of maintaining a good reputation (in terms of attracting new customers) is higher in credence goods markets than in ordinary and experience goods markets as choosing a credence goods provider involves more trust than choosing an ordinary goods or experience goods provider. While this explanation has some plausibility, more research on this issue is needed to identify the explanation for the differences in results. All in all, our results indicate that rating platforms contain valuable information for consumers in the market under consideration and that the dynamics are somewhat different compared to ordinary and experience goods markets.

However, while reviews are generally informative, this does not apply to all of them to an equal degree. Motivated by reports in the media about potentially fake or commissioned internet reviews,[42] 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 contrary to the negative association between ratings and prices for recommended reviews. That fake reviews make consumers more likely to choose lower quality products and that this leads to significant welfare losses is shown by Akesson et al. (2022) in a large online experiment with a nationally representative sample of respondents from the UK. The authors also suggest an interesting and easy to implement approach that seems promising to mitigate this problem: In their experiment, displaying warning banners about the possible presence of fake reviews (and advice on how to avoid being influenced by these reviews) reduces the adverse welfare effect of such reviews by 44 %.

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 specific 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."[43] The Yelp homepage states that the used recommendation software is entirely automated and regularly maintained and updated. It looks at hundreds of factors, including various measurements of quality, reliability, and user activity of reviewers in order to filter out unrecommended reviews. This process is dynamic and therefore, reviews that are recommended for any business can change over time as Yelp's software learns more about the reviewer and the business.[44] Our results suggest that the algorithm used by Yelp to discriminate between recommended and non-recommended reviews is helpful in identifying reviews that should not be trusted by consumers.[45] On the contrary, reviews on Google have less informational value for consumers, possibly because of their more lenient stance towards companies that are suspected of fake reviews. These results suggest that consumers in credence goods markets might benefit from taking the distinction between recommended and non-recommended ratings seriously, and by relying more on platforms that have more restrictive filters for potentially commissioned or fake reviews. From a policy perspective, this would imply that authorities could improve the outcome for customers in credence goods markets by obligating the platforms to provide a quality control of their reviews (or at least by informing the consumers of the potentially fake reviews). First attempts in this direction have been made by regulators in the UK who opened an investigation into whether Amazon and Google are doing enough to detect fake reviews, to investigate and promptly remove fake reviews, and to impose adequate sanctions on reviewers and businesses to deter them.[46]

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpubeco.2023.104891.

References

Abeler, J., Nosenzo, D., Raymond, C., 2019. Preferences for truth-telling. Econometrica 87, 1115–1153.
Akerlof, G.A., 1970. The market for "lemons": Quality uncertainty and the market mechanism. Quart. J. Econ. 84, 488–500.
Akesson, J., Hahn, R. W., Metcalfe, R. D., Monti-Nussbaum, M. (2022). The Impact of Fake Reviews on Demand and Welfare. https://conference.nber.org/conf_papers/f166391.pdf.
Anderson, E.T., Simester, D.I., 2014. Reviews without a purchase: Low ratings, loyal customers, and deception. J. Mark. Res. 51, 249–269.
Ashton, M.C., Kibeom, L., 2005. Honesty-Humility, the Big Five, and the Five-Factor Model. J. Pers. 73 (5), 1321–1354.
Balafoutas, L., Beck, A., Kerschbamer, R., Sutter, M., 2013. What drives taxi drivers? A field experiment on fraud in a market for credence goods. Rev. Econ. Stud. 80, 876–891.
Balafoutas, L., Kerschbamer, R., Sutter, M., 2017. Second-degree moral hazard in a real-world credence goods market. Econ. J. 127, 1–18.
Bindra, P.C., Kerschbamer, R., Neururer, D., Sutter, M., 2021. On the value of second opinions: A credence goods field experiment. Econ. Lett. 205, 109925.
Bohnet, I., Huck, S., 2004. Repetition and reputation: Implications for trust and trustworthiness when institutions change. Am. Econ. Rev. Pap. Proc. 94, 362–366.
Bolton, G., Katok, E., Ockenfels, A., 2004. How effective are electronic reputation mechanisms? An experimental investigation. Manag. Sci. 50, 1587–1602.
Bolton, G., Greiner, B., Ockenfels, A., 2013. Engineering trust – Reciprocity in the production of reputation information. Manag. Sci. 59, 265–285.
Bolton, G., Greiner, B., Ockenfels, A., 2018. Dispute resolution or escalation? The strategic gaming of feedback withdrawal options in online markets. Manage. Sci. 64, 4009–4031.
Cappelen, A., Halvorsen, T., Sørensen, E., Tungodden, B., 2017. Face-saving or fair minded: What motivates moral behavior. J. Eur. Econ. Assoc. 15, 540–557.
Chen, M.K., Chevalier, J.A., Rossi, P.E., Oehlsen, E., 2019. The value of flexible work: Evidence from Uber drivers. J. Polit. Econ. 127, 2735–2794.

[42] 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 14 December 2022).

[43] See https://www.nytimes.com/2011/05/22/your-money/22haggler.html (accessed on 14 December 2022).

[44] See https://trust.yelp.com/recommendation-software/ (accessed on 14 December 2022).

[45] For some shops, the distinction between recommended and non-recommended reviews can be strongly skewed, as the following shop with 6 recommended and 69 non-recommended reviews illustrates: https://www.yelp.com/biz/notebookservice030-berlin-3?osq=laptop+reparatur (accessed on 14 December 2022).

[46] See https://www.wsj.com/articles/amazon-google-probed-in-u-k-over-fake-reviews-11624615801 (accessed on 14 December 2022).

Darby, M., Karni, E., 1973. Free competition and the optimal amount of fraud. J. Law Econ. 16, 67–88.
Dulleck, U., Kerschbamer, R., 2006. On doctors, mechanics and computer specialists – The economics of credence goods. J. Econ. Lit. 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. Am. Econ. Rev. 101, 526–555.
Ely, J.C., Välimäki, J., 2003. Bad reputation. Q. J. Econ. 118, 785–814.
Ely, J.C., Fudenberg, D., Levine, D.K., 2008. When is reputation bad? Games Econ. Behav. 63, 498–526.
Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise - an experimental study on cheating. J. Eur. Econ. Assoc. 11, 525–547.
Gneezy, U., 2005. Deception: the role of consequences. Am. Econ. Rev. 95, 384–394.
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.
Gneezy, U., Kajackaite, A., Sobel, J., 2018. Lying aversion and the size of the lie. Am. Econ. Rev. 108 (2), 419–453.
Gottschalk, F., Mimra, W., Waibel, C., 2020. Health services as credence goods: A field experiment. Econ. J. 130, 1346–1383.
Grömping, U., 2007. Estimators of relative importance in linear regression based on variance decomposition. Am. Stat. 61, 139–147.
Grosskopf, B., Sarin, R., 2010. Is reputation good or bad? An experiment. Am. Econ. Rev. 100 (5), 2187–2204.
Hall, J., Kerschbamer, R., Neururer, D., Skoog, E. (2019). Uncovering sophisticated discrimination with the help of credence goods markups – evidence from a natural field experiment. Working Papers in Economics and Statistics 2019-11. University of Innsbruck.
Hillbig, B.E., Zettler, I., 2015. When the cat's away, some mice will play: A basic trait account of dishonest behavior. J. Res. Pers. 57, 72–88.
Hu, N., Bose, I., Koh, N.S., Liu, L., 2012. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decis. Support Syst. 52, 674–684.
Huck, S., Luenser, G., Tyran, J.-R., 2012. Competition fosters trust. Games Econ. Behav. 76 (1), 195–209.
Huck, S., Luenser, G., Spitzer, F., Tyran, J.-R., 2016a. Medical insurance and free choice of physician shape patient overtreatment. A laboratory experiment. J. Econ. Behav. Organiz. 131, 78–105.
Huck, S., Luenser, G., Tyran, J.-R., 2016b. Price competition and reputation in markets for experience goods. An experimental study. RAND J. Econ. 47, 99–117.
Iizuka, T., 2007. Experts' agency problems: Evidence from the prescription drug market in Japan. RAND J. Econ. 38, 844–862.
Kerschbamer, R., Sutter, M., 2017. The economics of credence goods – A survey of recent lab and field experiments. CESifo Econ. Stud. 63, 1–23.
Kocher, M., Schudy, S., Spantig, L., 2018. I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. Manag. Sci. 64, 3995–4008.
Luca, M., Zervas, G., 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. Manag. Sci. 62, 3412–3427.
Mayzlin, D., Dover, Y., Chevalier, J., 2014. Promotional reviews: An empirical investigation of online review manipulation. Am. Econ. Rev. 104, 2421–2455.
Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honestpeople: A theory of self-concept maintenance. J. Mark. Res. 45, 633–644.
Melnik, M.I., Alm, J., 2002. Does a seller's ecommerce reputation matter? Evidence from eBay Auctions. J. Ind. Econ. 50, 337–349.
Mimra, W., Rasch, A., Waibel, C. (2016a). Price competition and reputation in credence goods markets: Experimental Evidence. Games and Economic Behavior 100, 337-352.
Mimra, W., Rasch, A., Waibel, C. (2016b). Second opinions in markets for expert services: Experimental evidence. Journal of Economic Behavior and Organization 131, 106-125.
Ockenfels, A., Resnick, P., 2012. Negotiating reputations. In: Bolton, G., Croson, R. (Eds.), The Oxford Handbook of Economic Conflict Resolution. Oxford University Press, pp. 223–237.
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. Am. Econ. Rev. 92, 1093–1103.
Schneider, H.S., 2010. Moral hazard in leasing contracts: Evidence from the New York City taxi industry. J. Law Econ. 53, 783–805.
Schneider, H.S., 2012. Agency problems and reputation in expert services: Evidence from auto repair. J. Ind. Econ. 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).