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

DISCUSSION PAPER SERIES

IZA DP No. 15683

Reputation vs Selection Effects in Markets With Informational Asymmetries

Theodore Alysandratos
Sotiris Georganas
Matthias Sutter

NOVEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15683

Reputation vs Selection Effects in Markets With Informational Asymmetries

Theodore Alysandratos
Heidelberg University

Sotiris Georganas
City-University of London

Matthias Sutter
Max Planck Institute for Research on Collective Goods, University of Cologne, University of Innsbruck and IZA Bonn

NOVEMBER 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Reputation vs Selection Effects in Markets With Informational Asymmetries*

In markets with asymmetric information between sellers and buyers, feedback mechanisms are important to increase market efficiency and reduce the informational disadvantage of buyers. Feedback mechanisms might work because of self-selection of more trustworthy sellers into markets with such mechanisms or because of reputational concerns of sellers. In our field experiment, we can disentangle self-selection from reputation effects. Based on 476 taxi rides with four different types of taxis, we can show strong reputation effects on the prices and service quality of drivers, while there is practically no evidence of a self-selection effect. We discuss policy implications of our findings.

JEL Classification: C93, D82

Keywords: information asymmetries, reputation mechanisms, selection effects, credence goods, field experiment

Corresponding author:
Matthias Sutter
Max Planck Institute for Research on Collective Goods
Kurt-Schumacher-Strasse 10
53113 Bonn
Germany

E-mail: matthias.sutter@coll.mpg.de

* We thank Severine Toussaert, Sergiu Ungureanu as well as seminar participants in Lyon, MPI Bonn, Southern Denmark University, ESA 2018, EARIE 2018, TIBER 2018 and CRETE 2018 for useful comments, and Giulia Iori for her support. Stavrina Vlazaki, Venetia Nestoridou, Daphne Koletti and Stella Giapitzeli provided outstanding research assistance. Financial support from the Max Planck Society, the University of Cologne (through the Kelsen Prize) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/1– 390838866 is gratefully acknowledged.

1 Introduction

Informational asymmetries between buyers and sellers prevail in many markets and can, in the extreme, even lead to complete market breakdown (Akerlof, 1970). Expert professionals holding relevant information can cheat their less informed clients, which in turn leads to clients buying less services or leaving the market altogether. Examples of markets with asymmetric information abound, including legal services, financial advice, software programming, health care or repair services, and many more (Darby and Karni, 1973; Dulleck and Kerschbamer, 2006).

Modern technologies, such as rating platforms, promise to alleviate the problems of informational asymmetries. They allow for reputation-building such that trustworthy sellers can signal their qualities to buyers who then may refrain less from trading than without such reputation-building platforms. Many apps that match buyers and sellers rely on this approach to limit the negative effects of sellers' superior information on the likelihood of cheating buyers and on overall market efficiency. Yet, despite this, it is a challenge to identify whether such apps or platforms may improve overall efficiency, and, if so, through which channel. In fact, there are two potential mechanisms. First, the apps may indeed work because of their incentives to build up a good reputation. Second, the apps may be considered as working well because of self-selection. The latter means that more trustworthy sellers offer their services and products via app, while less trustworthy sellers sell their products without any devices that allow rating them. Depending upon which mechanism prevails, different welfare implications and policy conclusions arise. Yet, in a first step it is important to disentangle both channels, a challenge which is not easy to meet because it is required to hold one channel constant at a time while varying the other one.

In this paper, we present a field experiment that meets this condition. We ran our study in the taxi market in Athens, Greece, exploiting the simultaneous co-existence of various types of service providers that allow for a clean disentanglement of reputation and selection effects. We had research assistants take more than 400 taxi rides in Athens. Four RAs would take four different types of taxis at the same time for the exact same route, recording all

relevant variables (path taken, price, duration, service quality). Three of the types were regular yellow cabs that are driven by drivers who have passed a test and are officially accredited by the city of Athens to do their job. A noteworthy feature of Athens, however, is the fact that yellow cab drivers can not only be hailed on the street, but many of them are also generating rides through an app that is called *Beat*. So, *Beat*-drivers can be booked via the app or hailed on the street (in the which case they can still be identified by passengers as *Beat*-drivers via the app or via stickers on the car). When booked via the app, drivers get a rating from passengers; when hailed on the street, though, they can't get any rating because they were not booked via *Beat*. Any systematic difference between these two types of drivers (hailed on the street or booked via the app) can be attributed to reputation effects, since the selection effect is controlled for by only comparing drivers that have already been selected by *Beat* to work for them. In order to study the selection effect then, we compare unaffiliated yellow cabs (that are not registered on *Beat*) to *Beat*-drivers that were hailed in the street without the app. For both groups, reputation does not matter, because they can not be rated (and given that Athens has about 14,000 taxi drivers each taxi ride can practically be considered a single-shot game that rules out reputation-building). At the same time, unaffiliated yellow cab drivers differ from the *Beat*-drivers by the latter having self-selected for working on a platform that allows the rating of drivers. So, these three different types of taxi drivers allow for a clean identification of the reputation and selection effects of drivers into different suppliers of rides.

We add a fourth type of taxi, namely *Uber*-drivers. *Uber* works with its drivers outside the regulated yellow cab market and selects drivers solely based on its own criteria. Given that these drivers have no option to work as a yellow cab driver in case they are expelled from Uber (due to poor ratings) - while *Beat*-drivers have such an outside option if they missed the threshold rating that *Beat* requires - this feature suggests that reputation plays an even stronger rule for *Uber*-drivers than for *Beat*-drivers. *Uber* as the fourth type of taxi therefore allows to examine whether different degrees of reputational concerns - which are arguably the strongest for *Uber*-drivers, weaker for drivers booked by the *Beat*-app and weakest for yellow cab drivers and *Beat*-drivers hailed on the street without the app - lead

to different behavior and service quality of taxi drivers.

The results of our field experiment provide strong support for the reputation channel, while there is practically no evidence of self-selection going on. In terms of average price per trip, both yellow cabs and *Beat*-drivers who were hailed on the street - we call this condition BeatStreet from now on - were about 5-10% more expensive than the two types of drivers with reputational concerns, i.e., *Beat*-drivers who were booked via the app (referred to as BeatApp henceforth) and *Uber*-drivers. The latter charge the lowest prices of all when running pairwise comparisons. The higher prices in the first two conditions (yellow cab and BeatStreet) are due to a combination of factors: longer paths, wrong tariffs and (infrequently) more change kept by the driver. Driver ratings by our RAs (not on the platforms) offer an even starker picture. Regular yellow cab drivers are rated as 'good' or 'very good' about 35% of the time. This fraction is not much higher for BeatStreet-drivers (with 36%). Yet, BeatApp-drivers are rated as 'good' or 'very good' in 65% of cases, and *Uber*-drivers in 80% of the time. The service quality of the first two types of taxis is considered as much worse than the quality of the latter two types of taxis that can be rated by passengers. In particular, yellow cab- and BeatStreet-drivers are more often recorded as behaving or driving badly. So, overall we find no difference in prices charged and service quality of yellow cab and BeatStreet-drivers, which suggests that selection does not play a role in this market with asymmetric information between expert providers (the taxi drivers) and customers (their passengers). BeatApp-drivers perform much better in service quality and weakly better in prices, which indicates that reputation matters. Even more so, it matters most for *Uber*-drivers (who have no option of working as yellow cab- or *Beat*-driver), as they show clearly the best service quality. Their prices are also lowest on average, but that is probably due to *Uber*'s pricing algorithm, not because of the reputation channel of rating *Uber*-drivers.

The setting of our field experiment - a taxi market - is a common example of informational asymmetries between sellers and buyers. In such markets, feedback platforms have been shown to enable buyers to find more trustworthy sellers who have not exploited their informational advantage in the past (Bolton et al., 2004, 2013). So, it is known that such platforms matter for building up reputation and increasing market efficiency. Credence goods

markets–like expert services of lawyers, doctors, software programmers, repair specialists, or taxi drivers–are a prominent example of markets with asymmetric information, to the extent that buyers often cannot even judge ex post whether the type of service or good they have received is the one that would have been optimal (Dulleck and Kerschbamer, 2006; Balafoutas and Kerschbamer, 2020). In the first laboratory experiment on credence goods markets, Dulleck et al. (2011) have shown that reputation building significantly lowers overcharging of sellers. In the seminal field experiment in such markets, Schneider (2012) has compared the service quality of car mechanics if they encounter a customer only once vs. if repeat interaction with the same customer is possible. In the latter case, reputational concerns may matter, and Schneider (2012) finds, indeed, a small positive effect of this reputation channel in pricing. Similarly, Rasch and Waibel (2017) have reported that garages closer to highways tend to overcharge more, which they attribute to a higher likelihood of customer visits being just one-off instead of repeated. Kerschbamer et al. (2019) have found that repair shops with better ratings on internet platform charge on average lower prices for computer repairs. To our knowledge, none of these studies has been interested in disentangling reputation effects from self-selection effects with respect to apps that allow for the rating of sellers.[1]

Our paper is also related to the literature on the selection of professionals into specific branches of an economy. Most closely related are papers that study how social preferences relate to professional choices, because social preferences are also relevant for sellers' behavior in markets with asymmetric information (Kerschbamer et al., 2017). For example, Fisman et al. (2015) find that student subjects who focus on efficiency in experimental distribution games are more likely to choose employment in the private sector, while subjects who focus on equality are more likely to search for jobs in the non-profit sector. Friebel et al. (2019) compare behavior in an experimental trust game of police applicants (when they submit their application) and a sample of high school students in a similar age cohort. They find

[1]The first field experiment on the determinants of cheating in taxi rides was done by Balafoutas et al. (2013). They studied whether a taxi driver's overcharging and the extent of detours depend on a passenger's presumed familiarity with the city and the tariff system. Their finding was a resounding yes. Yet, their paper does not address reputation or self-selection effects in this market.

that the former group is more trusting and trustworthy than the latter group. Egan et al. (2019) provide evidence that there is self-selection of financial professionals with fraudulent records into companies with larger shares of employees with a similar history of misconduct. Gill et al. (2022) show that students who are less trustworthy in an experimental trust game are more likely to find their first job after graduation in the finance industry than students with higher levels of trustworthiness. In our paper, we are not interested in who is selecting to become a taxi driver (and what their personal or demographic characteristics might be). Rather, we want to study whether taxi drivers' self-selection into offering their services via an app with a rating system (*Beat*) matters for their provision behavior with respect to price and service when driving passengers from one place to another. So, we are interested in the selection into an environment where reputation can be built up, an issue which is of no concern to the papers discussed in this paragraph.

The remainder of the paper is organized as follows. Section 2 introduces the experimental design and our expectations. Section 3 describes the main results. Section 4 concludes.

2 Experimental design and expectations

For our field study, we hired four research assistants (blind to our research question and hypotheses) who were always simultaneously taking taxi rides from the same starting point to a particular destination. Together, they were asked to hail a yellow cab and a *Beat*-driver (BeatStreet), and book a *Beat*-driver (BeatApp) and a *Uber*-driver via the corresponding apps. Each route was then taken in four different taxis at practically the same time (to control for traffic and weather conditions). We call this set of four rides a quadruple.

We instructed our RAs to first use the Beat app to identify the location of Beat drivers in their close vicinity. One of them was then instructed to look for a *Beat*-driver in the streets (using this information but also looking at whether the driver had the app in use, or the company sticker on the side of the car), while another RA booked a *Beat*-driver via the app. A third RA hailed a yellow cab on the street, and a fourth RA booked a *Uber*-driver via the company's app. The order in which the RAs chose a particular taxi was changed

from quadruple to quadruple. In a few cases, it was difficult to complete a quadruple in the intended way, in which situation the remaining RA was asked to just take any other taxi to get to the destination (where the next quadruple would start). This explains why we don't have the exact same number of rides for each of the four taxi types.[2]

While in the car, the RAs were instructed to state their destination and always add the following statement: 'I have never been there, do you know where that is?'. This design choice corresponds to the "local-stranger" condition in Balafoutas et al. (2013), which was intended to make sure that drivers perceive the passenger to be less informed than they are (which is a necessary precondition to call the service a credence good). The assistants recorded the licence plate number (after arriving at the destination in order to control for multiple rides with the same drivers, which never happened), the estimated age of the driver and the start and end location. Since we also wanted to take account of service and driving quality, we asked the RAs to explicitly record occurrences of bad actions (crossing a red light, overtaking from the right, smoking in the car, smell of smoke in the car, texting, talking on a mobile phone, double hire, other) and of positive actionss (using a GPS, asking about a preferred route, asking about preferred radio stations, asking about car temperature, other). In addition to that, RAs had to rate the state of the car, the overall service provided by the driver and note down comments or any other extraordinary things that might happen.[3] Our RAs were all young (around 24) and female.

We chose a variety of routes, both in the center of the city and in the suburbs, including several metro stations, the main railway station, the port, a hotel, a well-known private university and a luxury shopping center (for a detailed list, see appendix A). The average route length was 10.7 km and duration 15.7 minutes, with 95% of the routes in the range of 2 to 21 km, respectively from 9 to 27 minutes. In total, we collected data for 476 rides, from 20 December 2017 until 28 March 2018, as Table I shows. This table also summarizes the characteristics of the four different taxi types and a few descriptive statistics about drivers.

[2]Looking at perfect quadruples, we have 93.

[3]A somewhat typical positive comment would be 'had bottles of water at every seat', or 'offered me candy'. A typical negative comment was 'nervous and bad driving' or 'bad smell'.

To form predictions about expected results, it seems straightforward to assume that drivers care primarily about their revenues through charging customers. Yet, those whose services are rated on a platform (i.e., those drivers that have been booked via *Uber* or the *Beat*-app) care also about staying with their rating above the threshold, the violation of which would lead to expulsion from the platform. To meet this criterion, drivers need to consider that passengers care about the price charged[4] and the service provided. Worse service and higher prices are likely to reduce the rating a driver will get from a passenger[5], for which reason we expect drivers who get rated to provide on average better service and lower prices (by taking less detours or avoid using wrong tariffs and add-on surcharges). While for BeatApp-drivers, the price is determined by the driver's choice, for *Uber*-drivers this is taken care of by the platform's algorithm - where relatively cheap prices are a means of building up a reputation for inexpensive rides. Regarding service quality, it can be expected that *Uber*-drivers provide the best service because their reputational concerns are the strongest among all taxis in a quadruple. This is the case because *Uber*-drivers have a lower outside option than BeatApp-drivers. The latter do have a licence to work as a regular yellow cab driver, but the former don't. For this reason, *Uber*-drivers have a lower continuation payoff in case of being expelled from the platform than a BeatApp-driver, which makes it more important for them to offer good service to get a good rating that keeps them above the required threshold. When comparing yellow cab drivers to BeatStreet-drivers–i.e., drivers hailed on the street without using the *Beat*-app, but who are registered on the app–their reputational concerns are very low in both cases, since both are not rated for that specific ride, for which reason we expect no difference in prices and services of these two types of drivers.

[4]Passengers may also care about the duration of a trip, yet time and price are highly correlated ($\rho = 0.56, p - value < 0.001$), so we ignore considerations of time here.

[5]Kerschbamer et al. (2019) find for computer repair shops an inverse relationship between prices charged and stars rated on rating platforms, which supports this expectation.

	Yellow	**BeatStreet**	**BeatApp**	**Uber**
Regulated yellow cab	yes	yes	yes	no
Hailed via	street	street	app	app
Reputation concerns	low	low	high	very high
Mean Driver age	52.3	50.4	48.5	40.1
Male	96.5%	97.7%	92.4%	90.8%
Nr of rides	114	126	117	119

Table I: Summary of the taxi types in the different treatments.

3 Results

3.1 Descriptives

We begin our presentation of results by providing descriptive statistics in Table II. In the first row, we show average prices, i.e., the fair paid by our RAs, contingent on the type of taxi taken. We see that *Uber* charges the lowest prices on average (9.9 Euro), followed by BeatApp (10.7 Euro). Yellow cabs charge on average 11.1 and BeatStreet 11.0 Euro. This ordering is compatible with our reasoning about what to expect. A Jonckheere-test for ordered alternatives (with yellow cab >= BeatStreet >= BeatApp >= *Uber*) provides $p < 0.01$. In pairwise comparisons, we find that *Uber* charges significantly lower prices than each of the other three taxi types (Wilcoxon test yields $p = 0.05$).

In the line below the average fare, we present in Table II the average experience rating by our RAs for the drivers. This rating ranged from 1 for very bad to 5 for very good. Yellow cabs and BeatStreet perform worst, with an average ranking of around 3.2. BeatApp performs already better, consistent with the effects of reputational concerns, with an average of 3.66 which is significantly better than Yellow cab and BeatStreet ($p < 0.05$ in both comparisons; Wilcoxon test). *Uber* performs best with an average rating of 3.99, which is better than any of the other ratings ($p < 0.05$ in each pairwise comparison; Wilcoxon test). A Jonckheere-test confirms the significant order yellow cab $\leq$ BeatStreet $\leq$ BeatApp $\leq$ *Uber* with $p < 0.01$. Looking at the relative frequency of ratings from 1 to 5 (in the middle part

	Yellow	**BeatStreet**	**BeatApp**	**Uber**
Mean Fare Paid	11.09	11.00	10.72	9.92
Mean Experience Rating (1:very bad)	3.16	3.24	3.64	3.94
Experience Rating (in percent):				
Very Good	2.63	3.15	6.78	16.67
Good	34.21	34.65	55.93	60.83
Average	43.86	47.24	32.2	22.5
Bad	14.91	13.39	5.08	0
Very bad	4.39	1.57	0	0
Mean Bad Actions	0.91	0.77	0.48	0.23
Mean Good Actions	0.39	0.55	0.77	0.93

Table II: Summary of Prices and Service Quality

of Table II), we note large differences in the extremes: Uber is never ranked as very bad or bad, but it is judged as very good in 19.5% of the rides, while yellow cabs are bad or very bad in 20.6% and very good in only 3.9% of the cases. The distribution for BeatStreet and BeatApp lies in between yellow cab and *Uber*.

At the bottom of Table II we present averages for bad actions and good actions of drivers with respect to driving and service quality. We add up the RAs recording of bad actions (crossing a red light, overtaking from the right, smoking in the car, smell of smoke in the car, texting, talking on a mobile phone, double hire, other bad action) and positive actions or gestures (using a GPS, asking about a preferred route, asking about preferred radio stations, asking about car temperature, other good action). We aggregate all good actions in one index and the bad ones in another, with equal weights (of 1) for all items, except for crossing red lights (weight=2) and double hiring (weight=2). The exception is motivated by the former capturing dangerous actions that might threaten the passenger's safety, and the latter being one of the main reasons why potential passengers might not enter a taxi even after it had stopped to pick them up. We had run a survey among 425 taxi customers in 2021 where about one fourth of survey participants (118 out of 425) indicated that it had happened

to them that they did not board a taxi when it had stopped for them, and they reported rudeness of the driver and other people being already in the taxi as the major reasons to do so.

As with the overall customer experience, the relative frequency of negative actions conforms with the expected ranking. There is a large and significant difference between Uber and BeatApp (Wilcoxon test yields $p < 0.01$), BeatApp and BeatStreet ($p < 0.01$), but not BeatStreet and Yellow ($p = 0.36$). In line with this observation, the median negative act is 0 for drivers booked through an app (either Beat or Uber), and 1 for divers hailed on the street. Regarding positive actions, the ranking is exactly the inverse. We find significant differences between all taxi types, except BeatStreet and Yellow ($p = 0.055$).

3.2 Regressions

Next we present several regressions that take account of the multiplicity of routes, the assistants' IDs, route length and duration and of the type of taxi used within a quadruple. In the first column of Table III we report the regression for the fare charged by a driver. As to be expected, this fare is larger if the distance and the trip are longer. The IDs for our RAs are insignificant, as they should be due to the randomizing of RAs into different rides in a quadruple. Among the dummies for the different types of taxis–yelow cab is taken as the benchmark and thus omitted–we see a significantly negative coefficient for *Uber*. BeatStreet and BeatApp are not significantly different from yellow cab. When comparing BeatStreet to BeatApp, we notice that prices in BeatApp are weakly significantly lower than in BeatStreet ($p < 0.1$).

In the second column of Table III, we look at service quality by regressing the passenger's experience rating on the variables already used in column 1. We note two assistants to be significantly negative, which possibly means they were slightly stricter in their evaluations. Trip distance and duration are not significant. Yet, experience ratings differ across taxi types. BeatApp and *Uber* are now both highly significant.[6] BeatApp trips are rated half a

[6]Restricting the sample to complete sessions (i.e. those where the RAs managed to find one of each four taxi types) and correcting the standard errors for clustering at the session level, yields the same significance.

	Fares	**Experience Rating**
Intercept	0.7	2.5
BeatStreet	-0.073	0.095
BeatApp	-0.297	0.469 **
Uber	-1.196 **	0.676**
Bench. Distance	0.828 **	0.001
Bench. Duration	0.09 **	0.008

Table III: Regression results. One star denotes significance at 5%, two stars at 1%. Controls for driver age/sex are included along with dummies for the assistants, but not presented in the table.

unit better than yellow cabs, and they are also better rated than BeatStreet rides ($p < 0.05$). *Uber*-rides are ranked best, with a coefficient of 0.77. Given that the scale ranges from 1 to 5, this coefficient indicates that rides with *Uber*-drivers are rated almost 20% better than rides with yellow cabs and BeatStreet.

Overall, the evidence on prices and service quality match our expectations fairly well, meaning that taxi rides where drivers are going to be rated (BeatApp and *Uber*) are on average cheaper and provide better service than rides where reputational concerns do not play a role. For self-selection, we do not see any obvious evidence, because we consistently find that yellow cab drivers and BeatStreet-drivers perform equally (bad).

4 Conclusion

Rating platforms persist in many different markets, covering, among others, holiday room bookings, professional expert services (e.g., medical, legal advice), software programming or repair shops. Such platforms are intended to improve market efficiency and alleviate informational asymmetries between sellers and buyers (Bolton et al., 2004, 2013). The potential effects of providing a service or selling a good over a platform may arise because of two effects: a selection effect–according to which different types of sellers self-select into the platform–and a reputation effect–which means that behavior of sellers changes in response

to their intention to build up a good reputation as a valuable means to attract also future customers (Grosskopf and Sarin, 2010; Huck et al., 2016). Disentangling these two effects to understand why rating platforms change the behavior of sellers is difficult because it requires holding one factor (either reputation or selection) constant while varying the other. We have exploited a unique setting which makes it possible, however, to distinguish between reputation and self-selection effect in a typical market with asymmetric information between buyers and sellers, namely the market for taxi rides.

More precisely, we have run our study in the taxi market in Athens, Greece, where we used the opportunity of different types of taxis being available at the same time. In addition to taking rides with traditional yellow cabs, we have used two types of taxis whose drivers are registered on the platform *Beat*, but who are at the same time certified (and in this capacity regulated) yellow cab drivers. One type of *Beat*-drivers was hailed on the street, in which case drivers could not be rated; the other type was booked via the app, which leads to a rating through the passenger. Comparing the latter two types of *Beat*-drivers reveals the immediate impact of a reputation building device, i.e., the reputation effect on drivers' pricing and service quality. Comparing regular yellow cabs with *Beat*-drivers hailed on the street allow examining the self-selection effect. We don't find any evidence for the latter–clearly indicating that also drivers booked via the *Beat*-app are not systematically different from yellow cab drivers. Yet, as soon as reputation kicks in, behavior of drivers gets noticeably more customer friendly. Drivers booked via the *Beat*-app charge (weakly) lower prices, but in particular they offer clearly better service and drive also more safely than those two taxi types without a rating opportunity (i.e., yellow cabs and BeatStreet). We consider these findings as strong evidence that rating platforms (at least in the taxi market) work mainly through the reputation effect, while self-selection effects seem to be negligible. By having also *Uber*-drivers in our set of taxis, we can show that even stronger reputational concerns–because *Uber*-drivers have no outside option of working as a yellow cab driver if they get expelled from the *Uber*-workforce–lead to even better service quality (and *Uber*'s pricing algorithm to the lowest prices on average). Again, this emphasizes the strong effect of the reputation channel on drivers' behavior.

Our results have important ramifications for policy making around the world. City administrations contemplating regulation against ride hailing apps have to include in their cost-benefit analysis the fact that there seems to be a substantial welfare increase when reputation platforms are in place, even if the same set of drivers that used to provide their service without an app would be shifting to providing it with a rating app. Apps do not just select the better drivers or, generally speaking, sellers (we see no evidence for this), but rather provide incentives for drivers to show their best side, a side they would not reveal when reputational concerns were absent. In this way, a reputation system does not seem to change the persons and their preferences, but it converts a game between sellers and buyers with no memory into a repeated game with memory, thus drastically changing the behavior of sellers on such markets.

References

[1] Akerlof, G. A. (1970) The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics* 84, 488-500.

[2] Balafoutas, L., Beck, A., Kerschbamer, R., Sutter, M. (2013) What Drives Taxi Drivers. A field experiment on fraud in a market for credence goods. *Review of Economic Studies* 80, 876-891.

[3] Balafoutas, L., Kerschbamer, R. (2020) Credence goods in the literature: What the past fifteen years have taught us about fraud, incentives, and the role of institutions. *Journal of Behavioral and Experimental Finance* 26, 100285.

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

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

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

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

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

[9] Egan, M., Matvos, G., Seru, A. (2019) The market for financial adviser misconduct. *Journal of Political Economy* 127, 233–295.

[10] Fisman, R., Jakiela, P., Kariv, S., Markovits, D. (2015) The distributional preference of an elite. *Science* 349(6254), aab0096.

[11] Friebel, G., Kosfeld, M. Thielmann, G. (2019) Trust the police? Self-selection of motivated agents into the German police force. *American Economic Journal: Microeconomics* 11, 59-78.

[12] Gill, A., Heinz, M., Schumacher, H., Sutter, M. (2022) Trustworthiness in the financial industry. *Management Science* forthcoming.

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

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

[15] Kerschbamer R., Neururer D., Sutter M. (2019) Credence Goods Markets and the Informational Value of New Media: A Natural Field Experiment. Working Papers in Economics and Statistics 2019-02. University of Innsbruck.

[16] Kerschbamer R., Sutter M., Dulleck U. (2017) How social preferences shape incentives in (experimental) markets for credence goods. *Economic Journal* 127, 393-416.

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

[18] Rasch, A., Waibel, C. (2018) What drives fraud in a credence goods market? Evidence from a field study. *Oxford Bulletin of Economics and Statistics* 80(3), 605-624.

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

[20] Wibral, M. (2015) Identity changes and the efficiency of reputation systems. *Experimental Economics* 18(3), 408-431.

A Appendix

A full list of the routes we used follows.

Agia Paraskevi - Maroussi

Acropolis Museum - Larissa Rail Station

Agia Paraskevi - Cholargos (pl Faneromenis)

Caravel - A Paraskevi, St. John

Caravel - Deree

Cholargos - Larissa station

Constitution - Glyfada, nymphon sq.

Glyfada - Acropolis Museum

Glyfada - Caravel

Golden Hall - Monastiraki

Larissa Station - Philadelphia Maroussi - Deree

Maroussi - Doukisis Plakentias station

Monastiraki - Glyfada

Monastiraki - Kolonaki

Monastiraki - Piraeus metro station

Philadelphia - Golden Hall

Piraeus - Acropolis Museum

Piraeus - Constitution

Syntagma - Piraeus (old station)

Larissa rail station - Fix

Doukisis Plakentias station - Cholargos, Palamas square