Disentangling Reputation from Selection Effects in Markets with Informational Asymmetries - A Field Experiment

Theodore Alysandratos*, Sotiris Georganas† and Matthias Sutter‡

Abstract

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. We show in a field experiment how to disentangle self-selection from reputation effects. Based on 476 taxi rides with four different types of taxis, we find strong evidence for reputation effects, but little support for self-selection effects. We discuss policy implications of our findings.

JEL classification: C93, D82.

*University of Heidelberg

†City, University of London, sotiris.georganas.1@city.ac.uk

‡Max Planck Institute for Research on Collective Goods, University of Cologne, University of Innsbruck, IZA Bonn and CESifo Munich

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 (Darby and Karni, 1973; Dulleck and Kerschbamer, 2006). The size of such markets is huge. Many of these markets are huge, like health care alone accounting for about 10% of global GDP[1], or financial and business services surpassing 20% of GDP in most rich countries.[2] At the same time, credence goods markets are plagued by fraudulent behavior, such as physicians providing unnecessary treatments (Gruber et al., 1999) or prescribing drugs with higher margins (Iizuka, 2007), or financial professionals cheating on their customers (Egan et al., 2019). Since informational asymmetries in credence goods markets can threaten market efficiency, it is important to understand how efficiency can be improved or restored.

Modern technologies, such as rating platforms, are a promising means 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. Yet, 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

[1]https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS

[2]https://w3.unece.org/PXWeb/en/CountryRanking?IndicatorCode=9

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 an 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, because in case of selection effects one might target entry qualifications and personal characteristics of sellers in particular markets, while in case of reputation effects the setup of feedback systems might be more important. Some governments try to protect taxi passengers from fraudulent behavior of taxi drivers (and thus from their personal characteristics) by enforcing a GPS-tracking module shown to passengers in order to protect fare-cheating. For example, the government of Thailand has recently passed a legislation that the country's 80.000 taxis must be equipped with cameras and a GPS system. The latter is explicitly intended to avoid fare-cheating (see https://www.nationthailand.com/in-focus/30319986). Similarly, in many countries of the world taxi drivers need to pass an exam to get a licence, which is another form of regulation to ensure entry qualification. Reputation systems are typically used in private companies like Uber and Lyft, for example, but much less so in publicly regulated taxi markets.

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. Research assistants took 476 taxi rides, split in sets of four rides each (called a quadruple henceforth) that were taken at the same time from the same origin to the same destination. Three rides in a quadruple were taken with regulated yellow cabs, yet two of them were also registered on an app that is called *Beat*. One of the *Beat*-drivers was hailed

from the street, and one via the app. The former cannot get a rating from a passenger, because that's only possible when hailed through the app. Therefore, we can compare the *Beat*-driver hailed on the street to the yellow cab driver in the quadruple that is not registered on *Beat*, and so we can estimate the effects of self-selection (of drivers who also work for *Beat*). The *Beat*-driver hailed via the app can be rated, and so reputational concerns become important. By comparing the two types of *Beat*-drivers - the one hailed via the app and the one hailed on the street - we can identify the reputation effect (because only the former can be rated, but both have self-selected into the app). The fourth driver in the quadruple was always from *Uber*. These drivers have no outside option of working as a regulated yellow cab driver, which makes reputational concerns particularly salient for them, as their rating has to meet a threshold in order not to be dismissed from *Uber*. So, *Uber*-drivers are included to examine whether their stronger reputational concerns lead to different service provision in comparison to *Beat*-drivers hailed via the app.

The results of our field experiment provide strong support for the reputation channel, while there is practically no evidence of self-selection going on. Prices are lowest for *Uber*-rides, where the company sets them. Yet, prices are also somewhat lower for *Beat*-drivers hailed through the app in comparison to the remaining two types of drivers in a quadruple. The quality of rides shows an even stronger pattern. Our research assistants (RA's) rated *Uber*-rides clearly best, followed by app-generated rides with a *Beat*-driver. The poorest ratings were for *Beat*-drivers hailed on the street and for the regular yellow cab drivers. Note that the latter two types of drivers could not be rated, for which reason reputational concerns cannot matter. Yet, self-selection does not matter as well, as *Beat*-drivers hailed on the street perform equally poorly as regular yellow cab drivers that did not register on *Beat*.

Our paper contributes to the literature on informational asymmetries between buyers and sellers on markets and how reputation-building devices can help improve efficiency 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; Bohnet and Huck, 2004; Huck et al., 2016).[3] So, it is in general known that such platforms matter for building up reputation and increasing market efficiency. Yet, credence goods markets are particularly prone to incentives for cheating because buyers cannot even judge *after* purchasing a good or service whether that was what they actually needed (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. A similar result is reported in Rasch and Waibel (2018) who report 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. The evidence in Mimra et al. (2016) is less clear, suggesting that reputation systems need not necessarily reduce the level of fraud. Kerschbamer et al. (2023) have found that repair shops with better ratings on internet platform charge on average lower prices for

[3]This is even true when sellers can change their identity by creating a new one to shed a negative old identity, as Wibral (2015) shows. Even in such a situation, trustworthiness is higher than in the complete absence of a reputation system (but lower in comparison to a situation where identities cannot be changed.

computer repairs, suggesting a positive effect of reputational concerns on the provision behavior of sellers. Our current paper is novel because none of the studies mentioned so far has disentangled reputation effects from self-selection effects.

A paper by Liu et al. (2021) is most closely related to ours. They compare the driving behavior of Uber and Taxi drivers in New York City. They find that for short trips, driving distances are very similar within matched Taxi-Uber pairs, but on airport trips, Taxi rides are longer. While they can rule out drivers selecting into specific routes, they cannot completely rule out the possibility that the underlying distributions of honesty-types of Taxi and Uber drivers differ. Our comparison between regular yellow cab drivers and *Beat*-drivers hailed on the street accounts for selection into the app, whereas the comparison between the two types of *Beat*-drivers (hailed on the street vs. via the app) accounts for the effect of reputation. Liu et al. (2021) also show that both the pricing scheme and the set of technological mechanisms decrease moral hazard. However, as the authors admit, they cannot evaluate one independently of the other. Since all *Beat*-drivers are subject to the same pricing scheme as other cabs, we are evaluating the effect of the technological mechanism that rewards reputation alone. Contrary to Liu et al. (2021), we also consider the quality dimension of taxi rides, finding large increases in the quality of the services provided by drivers with reputational incentives. This indicates that even if price differences are small, consumer welfare is increased under a reputation mechanism thanks to incentives to compete in the quality dimension.

2 Experimental design

Treatments 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. We call a set of four rides a quadruple. Within a quadruple, we had four types of taxi rides, characterized as follows: *Yellow* rides are in regular yellow cabs. Their drivers are officially accredited by the city of Athens to do their job, and their fares are regulated by the city with respect to charges per kilometer or waiting times. *BeatStreet*-rides are also provided by accredited taxi drivers, but these drivers are also registered on the app *Beat* that generates matches between customers and drivers who are registered on the app. For the latter, a driver must meet a certain threshold (of 4.5) in ratings from passengers. Importantly, *BeatStreet*-rides were hailed on the street, *not* via the app (but *Beat*-drivers can still be identified via the app or through a sticker in the car). The third type of taxi ride in a quadruple we call *BeatApp*. These are drivers hailed via the *Beat*-app, which means that they get a rating from the passenger, which is *not* the case in *BeatStreet*.[4] Any systematic difference between *BeatStreet* and *BeatApp* 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 *Yellow*-drivers (who are not registered on *Beat*) to *BeatStreet*. 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). Finally, as a fourth

[4]Note that it never happened that a driver asked one of our under-cover passengers to cancel the reservation via the app (in order not to get rated on the platform).

type of ride we included a ride with *Uber*. *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 (which happens in case of 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 lead to different behavior and service quality of taxi drivers. Such concerns are arguably the strongest in *Uber*[5], weaker for *BeatApp*, and weakest for *BeatStreet* and *Yellow*.

Implementation: 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 (*Yellow*), 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.[6]

[5] *Uber* drivers have a strong incentive to retain sufficiently high ratings in order not to be expelled from the platform. This means that *Uber*-drivers with very poor quality are likely to be expelled, leaving the more customer-friendly drivers in their pool.

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

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 actions (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.[7] 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 1 shows. This table also summarizes the

[7]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'.

characteristics of the four different taxi types and a few descriptive statistics about drivers.

Table 1: Summary of the taxi types in the different treatments

	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

Expectations: 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 (*Uber* and *BeatApp*) care also about staying with their rating above the threshold (of 4.5 on both platforms), 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[8] and the service provided. Worse service and higher prices are likely to reduce the rating a driver will get from a passenger[9], for which reason we expect drivers who get rated to provide on average better service and lower prices (this is possible in *BeatApp* by taking less detours or avoid using wrong tariffs and adding unwarranted surcharges; in *Uber* the price is set by the company and therefore not under control for the driver).

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 drivers on *BeatApp*-rides. 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 the former

[8]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.

[9]Kerschbamer et al. (2023) find for computer repair shops an inverse relationship between prices charged and stars rated on rating platforms, which supports this expectation.

to offer good service to get a good rating that keeps them above the required threshold.[10] Regarding *Yellow*-drivers and those on *BeatStreet*, the reputational concerns are very low for both drivers, since both are not rated for that specific ride, for which reason we expect no difference between both types of drivers.

3 Results

3.1 Descriptives

We begin our presentation of results by providing descriptive statistics in Table 2. In the first line, we show average prices, i.e., the fare paid by our RAs, contingent on the type of taxi taken. The average prices are descending from left (*Yellow*) to right (*Uber*), which is compatible with our reasoning about what to expect. A Jonckheere-test for ordered alternatives (with $Yellow >= BeatStreet >= BeatApp >= Uber$) yields $p < 0.01$. In pairwise comparisons, we find that *Uber* charges significantly lower prices than each of the other three taxi types (Wilcoxon test yields $p = 0.05$). While pairwise tests reveal no significant differences for comparisons between Yellow, BeatStreet and BeatApp, the regressions in the next subsection will also reveal a weak difference between *BeatApp* and *BeatStreet*.

In the line below the average fare, we present in Table 2 the average experience rating by our RAs. This rating ranged from 1 for very bad to 5 for very good. *Yellow* and *BeatStreet* perform worst, with an average ranking of around 3.2 (with no significant difference between both types of rides). *BeatApp* performs already better, consistent with the effects

[10] A survey we ran among taxi customers (to be discussed in more detail later) confirms that they do care about service quality and that the #1-reason for giving ratings is to inform the platform about a driver's quality.

of reputational concerns, with an average of 3.65, which is significantly better than *Yellow* and *BeatStreet* ($p < 0.05$ in both comparisons; Wilcoxon test). *Uber* performs best with an average rating of 3.94, 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* $\leq$ *BeatStreet* $\leq$ *BeatApp* $\leq$ *Uber* with $p < 0.01$.

Looking at the relative frequency of ratings from 1 to 5 (in the middle part of Table 2), we note large differences in the extremes: *Uber* is never ranked as very bad or bad, but it is judged as very good in 16.81% of the rides, while *Yellow*-rides are rated as bad or very bad in 19.3% and very good in only 2.63% of the cases. The distributions for *BeatStreet* and *BeatApp* lie in between *Yellow* and *Uber*, but only ratings in *BeatApp* are clearly better than in *Yellow* (even though any *Beat*-driver is also a regulated and accredited yellow cab driver).

At the bottom of Table 2 we present averages for bad actions and good actions 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. In this survey, we had also asked respondents about their assessment of the drivers and the cars of three different types of taxis, i.e., yellow cab, *Beat* and *Uber*. The rating of yellow cabs and their drivers was always the worst one, while *Beat* and *Uber* were roughly rated equally.

Turning back to the results from our field experiment, we see at the bottom of Table 2 the results on bad and good actions of drivers in the different types of taxis. 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 between *BeatStreet* and *Yellow* ($p = 0.36$). Regarding positive actions, the ranking is exactly the inverse. We find significant differences in all pairwise comparisons between the different taxi types, except for the comparison between *BeatStreet* and *Yellow* ($p = 0.055$).[11]

[11]Note from Table 2 that *Uber*-drivers had an average score of 0.93 for good actions. We counted using a GPS as a good action, and since *Uber* requires drivers to use GPS, they should have had a minimum score for good actions of 1.00. Yet, our RAs informed us that many *Uber*-drivers did not use the GPS (at least not visibly on their cell-phone). If we excluded the use of a GPS as a good action, the averages for good actions would be 0.32 for *Yellow*, 0.35 for *BeatStreet*, 0.49 for *BeatApp*, and 0.47 for *Uber*.

Table 2: Summary of Prices and Service Quality

	Yellow	**BeatStreet**	**BeatApp**	**Uber**
Mean Fare Paid (in Euro)	11.09	11.00	10.72	9.92
	(0.4)	(0.41)	(0.37)	(0.33)
Mean Experience Rating (1:very bad)	3.16	3.25	3.65	3.94
	(0.081)	(0.07)	(0.063)	(0.057)
Experience Rating (in percent):				
Very Good	2.63	3.16	6.84	16.81
Good	34.21	34.92	56.41	60.5
Average	43.86	46.83	31.62	22.69
Bad	14.91	13.49	5.13	0
Very bad	4.39	1.59	0	0
Mean Bad Actions	0.91	0.77	0.48	0.23
	(0.1)	(0.09)	(0.08)	(0.06)
Mean Good Actions	0.39	0.55	0.77	0.93
	(0.05)	(0.06)	(0.06)	(0.06)

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 3 we report the regression for the fare charged by a driver. As expected, this fare is larger if the distance and the trip's duration are longer (according to a benchmark distance and route; see the legend to the table). The IDs for our RAs are insignificant, as they should be due to randomizing RAs into different rides in a quadruple. Among the dummies for the different types of taxis – *Yellow* is taken as the benchmark and thus omitted – we see a significantly negative coefficient for *Uber*, meaning that *Uber*'s pricing algorithm (which is not in the hands of *Uber*-drivers) is significantly cheaper than what the drivers of the regulated taxis charge. *BeatStreet* and *BeatApp* are not significantly different from *Yellow*. When comparing *BeatStreet* to *BeatApp*, however, we notice that prices in *BeatApp* are weakly significantly lower than in *BeatStreet* ($p < 0.1$), suggesting a weak effect of reputational concerns on prices. We also control for the estimated age of drivers, which seems to reduce prices slightly, and driver's sex (being male is related to slightly higher prices). For the latter variable, note that there is hardly any variance, however (see Table 1).

In the second column of Table 3, we look at service quality by regressing the passenger's experience rating on the variables already used in the first column. Experience ratings differ across taxi types. *BeatApp* and *Uber* are both highly significant.[12] *BeatApp* trips are rated

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

half a unit better than *Yellow*, and they are also better rated than *BeatStreet* ($p < 0.05$). *Uber* is ranked best, with a coefficient of 0.676. Not shown in the table, we note that two assistants have a significantly negative dummy, which possibly means they were slightly stricter in their evaluations. Trip distance and duration are not significant, but older drivers seem to provide worse quality, and so do men.

Overall, the evidence on prices and service quality matches our expectations fairly well, meaning that taxi rides where drivers are going to be rated (*BeatApp* and *Uber*) provide better service than rides where reputational concerns do not play a role. Prices are cheapest in *Uber*, where the company's algorithm determines the fare. This can be interpreted as an institutional reputation, rather than a personal one (of a driver). For *BeatApp* we find weakly significantly lower prices than for *BeatStreet*, which provides some evidence that also personal reputation may affect prices (by taking less detours if a driver is rated). Taking both service quality and prices into account, we observe evidence in favor of reputation effects. On the contrary, given the persistent null-results when comparing *Yellow* and *BeatStreet*, we do not observe any evidence of selection effects.

Table 3: Regression results.

	Fares	**Experience Rating**
Intercept	0.7	3.5
	(0.82)	(0.26)
BeatStreet	-0.073	0.095
	(0.29)	(0.095)
BeatApp	-0.297	0.469 **
	(0.3)	(0.096)
Uber	-1.196 **	0.676**
	(0.32)	(0.1)
Est. Driver Age	-0.0015 **	-0.0067941**
	(0.0099512)	(0.0032136)
Driver Sex (1 = men)	0.0042 **	-0.09665**
	(0.44995)	(0.1453)
Benchmark Distance	0.828 **	0.001
	0.05	0.01
Benchmark Duration	0.09 **	0.008
	0.03	0.01
R^2	0.716	0.218
Adjusted R^2	0.709	0.2
Observations	476	476

One star denotes significance at 5%, two stars at 1%. The benchmark durations and distances were measured for every route using google maps on the same weekday at 5am, without traffic. Controls for driver age/sex are included along with dummies for the assistants, but not presented in the table. Dummies for assistants are insignificant for fares, but significant for two assistants for experience rating.

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 (Balafoutas and Kerschbamer, 2020; Kerschbamer et al., 2023). 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 to distinguish between reputation and self-selection effects 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 (*Yellow*) with *Beat*-drivers hailed on the street allows 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. This is clearly reflected in better service quality (as measured in the experience rating), and partly also in slightly cheaper prices in *BeatApp* than in *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 and they also shed light on attempts to shut down or forbid ride hailing apps such as *Uber* or *Beat*. A few months after our field experiment, the city of Athens had forbidden *Uber* to offer its services in town. The same had happened in other cities, like for example Vienna, Austria. City administrations contemplating regulation against ride hailing apps should 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 (as in the case of *Beat*). 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.

Acknowledgements

We thank Raymond Fisman, four referees, Severine Toussaert, Sergiu Ungureanu as well as seminar participants in Lyon, MPI Bonn, Southern Denmark University, Southwestern University of Finance and Economics, ESA, EARIE, TIBER and CRETE for useful comments, and Giulia Iori for her support. Stavrina Vlazaki, Venetia Nestoridou, Daphne Koletti and Stella Giapitzeli provided outstanding research assistance. An anonymous Beat ex executive provided very useful comments from the start. 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.

References

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

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

[3] Balafoutas, L., Kerschbamer, R., "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 (2020), 100285.

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

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

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

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

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

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

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

[11] Gruber, J., Kim, J., Mayzlin, M., "Physician fees and procedure intensity: The case of Cesarean delivery." *Journal of Health Economics* 18 (1999), 473-490.

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

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

[14] Kerschbamer R., Neururer D., Sutter M., "Credence goods markets, online information and repair prices: A natural field experiment." *Journal of Public Economics* 222 (2023), 104891.

[15] Liu, M., Brynjolfsson, E., Dowlatabadi, J., "Do digital platforms reduce moral hazard? The case of Uber and taxis." *Management Science* 67 (2021), 4665-4685.

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

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

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