Uncertainty and Reputation Effects in Credence Goods Markets

Eric Schniter[*], J. Dustin Tracy[†], and Vojtěch Zíka[‡]

August 30, 2021

Abstract

Credence-goods experiments have focused on stylized settings in which experts can perfectly identify the buyer's best option and that option works without fail. However, in nature credence goods involve uncertainties that complicate assessing the quality of service and advice. We introduce two sources of uncertainty into a credence goods experiment. The first is diagnostic uncertainty; experts receive a noisy signal of buyer type so might make an 'honest' mistake when advising what is in buyers' best interests. The second is service uncertainty; the services available to the buyer do not always work. Both sources of uncertainty make detection of expert dishonesty more difficult, so are expected to increase dishonesty by experts and decrease buyer trust (willingness to consult experts for advice and to follow expert advice) – decreasing efficiency of the interactions. We also analyze how buyers use ratings and whether ratings restrain both dishonesty and distrust by creating reliable reputations. In contrast to predictions, we find that uncertainty decreases dishonesty and increases trust. Also in contrast to predictions, ratings do not improve efficiency of the transactions under uncertainty – in part due to buyers' tendency to 'shoot the messenger' (give low ratings) when they buy service that does not work due to bad luck, and to give experts the 'benefit of the doubt' (high ratings) when they buy service that may have been intentionally over-provided (not in the buyer's best interest).

Keywords: Credence Goods, Uncertainty, Principal Agent, Ratings, Experiment

JEL-Codes: *D82, L14, L15*

Recommended Citation:
Schniter, E., Tracy, J. D., & Zíka, V. (2021). Uncertainty and reputation effects in credence goods markets. ESI Working Paper 21-15. `https://digitalcommons.chapman.edu/esi_working_papers/352/`

We thank Prithvijit Mukherjee, David Rojo Arjona, Marco Schwartz and session participants at: the 2021 Annual Meeting of the Austrian Economic Association, the 2020 Southern Economic Association Annual Meeting, and 2019 Economic Science Association North American Meeting for comments and feedback.

[*]Economics Science Institute, Chapman University & Division of Anthropology, California State University, Fullerton; eschniter@fullerton.edu

[†]Economics Science Institute, Chapman University; tracy@chapman.edu

[‡]Economics Science Institute, Chapman University & Jan Evangelista Purkyně University; zika@chapman.edu

1 Introduction

Trust and honesty are essential for markets to function; if consumers[1] do not trust, they will not enter the market, forgoing mutually beneficial exchanges. If providers of goods and services are not honest, there is also an efficiency loss because exchanges occur at a net-loss to the consumers, which is usually larger than the providers' gain. This might undermine consumers' trust and deter future exchange. For many goods and services, provider reputation can ensure high efficiency. Consumers can easily ascertain if the good or service meets their needs. Any dishonest provider is quickly identified and ostracized. However, in other situations, it is difficult to judge whether the recommendation we receive from a provider is honest. For example, we rely on doctors for advice about the best medical treatment. We might know after the advised treatment that our health improved. Yet, what we do not know is whether a less expensive treatment could have achieved the same result. Sometimes, it is apparent that the advised service did not achieve the desired outcome; for instance, after the medical treatment we remain ill. A host of other situations force consumers who rely on expert advice to select the best product or service despite the difficulty they may have judging the quality of the advice. The goods and services in such situations, termed *credence goods*, include car repairs, financial advice, and taxi rides in unfamiliar places.[2]

Darby and Karni (1973) coined the term credence goods. Credence goods research using experiments stemming from Dulleck et al. (2011) is summarized by Kerschbamer and Sutter (2017). However, as Kerschbamer and Sutter (2017) note, experiments have focused on situations in which uncertainty is not present. By starting with experiments investigating decisions under certainty, researchers gained insights into the dynamics of these markets. The literature shows that reputation and liability (recourse when the service is not adequate) increase honesty and trust, when there is certainty. However, these markets rarely are characterized by simple, certain solutions. This is one reason we rely on experts' advice and services. Unfortunately, even honest advice from an expert can lead to undesirable outcomes.

We contribute to the credence goods literature through an experiment with two sources of uncertainty: diagnostic uncertainty and service uncertainty. Diagnostic uncertainty is an instance in which an expert gives advice based on information that might not be correct, albeit unintentionally and through no fault of their own; e.g., a doctor makes a diagnosis based on an test that is prone to inaccuracy. Service uncertainty is an instance in which the most appropriate service is not guaranteed to work; e.g., a doctor recommends a treatment that is known to work only 67% of the time but nonetheless has a net positive expected benefit. This uncertainty creates plausible deniability (Gillies and Rigdon, 2019) for any expert who provides dishonest advice, which in turn limits the disciplining effect of reputation. Rubin and Sheremeta (2016) find that the introduction of a productivity shock in a principal-agent experiment reduces both wages and effort. Balafoutas et al. (2020) in a paper contemporaneous to ours, also report on experiments with diagnostic uncertainty, including a treatment in which the accuracy of the diagnosis is endogenous.

We make an additional contribution to the literature by implementing a reputation mechanism, based on those popular in an exchange market outside the laboratory. Reputation treatments in laboratory experiments have used fixed buyer-expert pairs: buyers are assigned to an expert only once and then remain in those pairs. Our design randomly reassigns buyers and experts after each interaction. However, buyers rate the experts they consult and those ratings are available to future buyers. We allow three ratings: "satisfied", "neutral" or "unsatisfied", which is much like the rating system on eBay. To our knowledge, we are the first credence good experiment to use a rating system. In theory, if a reputation system worked perfectly and conveyed an accurate summary of the history, a buyer who saw any rating would make the same decisions about interactions with the rated expert as would the buyer who made the rating, then ratings are equivalent to fixed pairs. If the rating system is noisy or used imperfectly, as we expect it will be, the results would be noisier than results from an experiment with fixed pairs. Bohnet and Huck (2004) found in a trust game that subjects in the "reputation-stranger" treatment more often trusted and were trustworthy than subjects in the stranger treatment and less often than in the fixed-pairs treatment. It is also possible that experts may try to exploit the noise. Dishonest behavior will increase because there is a (perceived) decrease in the

[1]In this section, we have changed our terminology to "consumer" and "provider", terms typical to discussion of trade and markets, rather than "expert" and "buyer", terms commonly used in credence goods research.

[2]Ride-sharing apps correct for asymmetric information, so with them taxi-rides might now be removed from the list.

likelihood that the dishonesty will be communicated.

We report results from an online experiment, in which buyers and experts both participate for three rounds. In each round, buyers were randomly assigned to a need type, either high or low, but were not informed of their type. If consulted, experts would run a test that provided a signal of the buyer's type. Experts chose what advice to give for each signal they could receive. Two types of service were available; one was best for high-need buyers and one was best for low-need buyers. The service best for high-need buyers earned the expert more profit; thus, buyer and expert interests were aligned. However, low-need buyers created conflicted motives for the expert, who could either give honest advice or earn more immediate profit by giving dishonest advice that the buyer follows. We ran six cells of a 2x2x2 design, in which we varied whether experts had reputations and each source of uncertainty, but not both types of uncertainty together. In the Diagnostic Uncertainty treatment, the expert received a correct signal about the buyer's type only 3/4 of the time (rather than always). In the service uncertainty treatment, the service only worked 2/3 of the time. In the reputation treatments, after the buyers were informed of their payment amounts, they rated their satisfaction with the advice the expert provided using the ratings "satisfied", "neutral" or "unsatisfied". Ratings were made available to buyers who encountered that expert in later rounds. In the second and third rounds, buyers had the option of not consulting the expert, but then only received the service best for low-need types. Ratings provide buyers with the potential to impact the expert's future profits.

We found, despite the cover provided by uncertainty, which can make dishonesty more profitable, experts were no more likely to provide dishonest advice to low-need type buyers. Consistent with previous studies, we found, without uncertainty, experts were more likely to provide dishonest advice when there were no ratings. Surprisingly, when there was uncertainty, rates of dishonest advice were no higher in the treatment cells where consumer did not rate experts. Equally surprising is that the lack of ratings did not create any buyer hesitation; in the first round, before any reputation was formed, there were no statistically significant differences in advice following. Consultation of experts in the uncertainty conditions had increased, despite that their advice was less informative. In the diagnostic uncertainty conditions, buyers seemed to give experts the benefit of the doubt; satisfied rating were the most common rating when low-need buyers selected high-need service. In the service uncertainty treatments, when the service did not work, buyers tended to "shoot the messenger" and give the unsatisfied rating for the high-need service but not for the low-need service. Like other unsatisfactory ratings, they decreases consultation, which led to efficiency loss. Our results suggest that ratings do not improve outcomes and may worsen them when there is uncertainty.

The paper proceeds as follows. The next section reviews the related literature and forms hypotheses. The third section present our design. The fourth section presents our results. The final section discusses results and relates them to the literature.

2 Background

While issues arising from exchanges in which both parties do not have the same information have been recognized for centuries (Rowell and Connelly, 2012), Arrow (1963) and Akerlof (1970) are pivotal in theoretical work on asymmetric information. Arrow (1963) in particular, calls attention to doctors' conflicting interests between prescribing treatment that is best for the patients' health and that is most profitable.

Much work has been done to formalize and expand the theory on asymmetric information. Darby and Karni (1973, p. 69) define credence goods to have value which "cannot be evaluated in normal use. Instead the assessment of their value requires additional costly information." In contrast, experience goods may entail information asymmetry before they are experienced, after normal use there is no information asymmetry. We may have to take the waiter's advice about the quality of a dish, but as soon as we taste it we know the true quality. Plott and Wilde (1980) present a model in which consumers facing experts with a conflict of interest search until search cost exceeds expected benefit of further search. Crawford and Sobel (1982) show that under information asymmetry, unless interests align, the better informed party will introduce self-serving noise. The signal sent maximizes the sender's (expert's) expected profit, balancing the gains if the receiver (buyer) trusts the signal against the costs from actions (not) taken when a signal is not trusted. In contrast to previous models, in which need level could take on any value within a given range, Pitchik and Schotter (1987) present a discrete model in which need level is either high or low. Most experiments use this binary model. They also argue an increase in expert certainty, will lead to decrease in honesty, though this depends on heterogeneity of expert certainty. Wolinsky (1993) posits a model in which expert diagnosis is imperfect. However, in contrast to how we model uncertainty, in Wolinsky's model, experts present customers an estimate (of problem size and thus price). Misdiagnosis does not lead to un-repaired or over-repaired problems, just surprise bills. Wolinsky's model also assumes the benefit of service is always great enough to ensure the customer enters the market. While discussing "reputation" what Wolinsky (and many others) models is personal history with an expert, not any capacity to share an opinion about an expert to acquaintances. Bester and Dahm (2018) assume diagnosis requires costly and unobservable effort, which contributing to inefficacy. They argue that accurate buyer reports of the failure of treatment can induce expert honesty, if reports of misdiagnosis imply additional cost to experts. Liu et al. (2019) present a model where expert ability is heterogeneous, and low ability experts make diagnostic errors but high ability experts do not.

The credence goods literature has identified three ways experts can be dishonest: *underprovision*, in which a high-need buyer is only provided the low-need remedy; *overcharging*, in which the buyer is billed for and allegedly provided the high-need remedy but in reality only provided with the low-need remedy; and *overprovision*, in which a low-need buyer is provided the high-need remedy (Dulleck and Kerschbamer, 2006).

Our review of the related experimental literature, starts with papers not always cited in credence good reviews. Plott and Wilde (1982) present evidence from experiments that when buyers have uncertainty regarding their need sellers will advise buyers to purchase options that increase the sellers' profits. Gneezy (2005) finds evidence of aversion to lying in a sender receiver game; compared to allocations in a binary dictator game, subjects were less likely to send a deceptive message likely to result in the allocation. Rates of anticipated and actual advice following were both ∼80%. As expected, lying was more likely when it was more profitable. In a 2 x 2 design, Sánchez-Pagés and Vorsatz (2007) vary both the profitability of lying and the option of a costly punishment in a sender receiver game with repeated rounds and random matching. They find that punishment does not have a statistically significant impact the likelihood of lying but does increase the likelihood that the receiver 'trusts' the message. Rates of lying did not statistically differ depending on how profitable it was. As expected, subjects punished lying, particularly when they trusted the lie. There were also fairly frequent rates of punishment when the message was truthful but not trusted, 5% and 13% depending on profitability of lying, that the authors attributed to subject error.

Laboratory experiments have identified important factors in restraining expert dishonesty. Dulleck et al. (2011) report on a credence goods experiment with 16 treatment cells that vary whether there is reputation (if seller has an identity rather than being anonymous), competition (if the buyer can choose from multiple sellers), liability (if the buyer has recourse when undertreated), and verifiability (if overcharging is possible). They find evidence of overcharging, overtreatment, and undertreatment, concluding that verifiability does little to improve efficiency but liability increases efficiency. Reputation has little impact and competition reduces prices. Kerschbamer et al. (2017) present evidence that heterogeneity of social preferences explains why verifiability does not increase efficiency. Beck et al. (2014) find that, compared to traditional undergraduate students, students training to be car mechanics are more dishonest in a laboratory credence goods experiment. Bejarano et al. (2017) find that subjects who select into payment schemes with a conflict of interest[3] exhibit more dishonesty as experts.

Field experiments expand our insight and show that many of the lab finding generalize. In a pair of field experiments focused on auto mechanics, Schneider (2012) finds significant levels of both over and undertreatment, and that the suggestion of repeated business did not significantly improve recommendations. We note that despite the undertreatment, the mechanics did not leave much money on the table. In a field experiment, Balafoutas et al. (2013) find that when the customer is perceived to be not from the city or country (and less familiar with what route should be taken) Athens taxi drivers choose a longer, more expensive route, and were more likely to overcharge the customer. In a related field experiment, Balafoutas

[3]A conflict of interest arises when the expert makes greater profits from particular (credence) goods. These are the case most often studied but not inherent.

et al. (2017) find that "second-degree moral hazard" situations in which the buyer will be reimbursed for the charges, also increase overcharging. Gottschalk et al. (2020) find that 28% of Swiss Dentists recommended unnecessary fillings; lower-income of patients, shorter waiting times for appointments, ownership of the practice were associated with increased likelihood of overtreatment.

While reputation in the context of credence goods is generally induced through repeated interactions and personal history, there is considerable research showing that consumer ratings of sellers can identify low-quality products and discipline dishonest sellers. Cabral and Hortaçsu (2010) find that negative feedback on eBay can have a deleterious impact on a sellers' future sales. Luca (2016) finds that one star increase in a seller's average Yelp review is associated with a 9% increase in revenue. Reimers and Waldfogel (2021) show that consumer reviews increase total sales and improve consumer welfare. Positive reviews have more impact than negative reviews suggesting that the availability of reviews increases market participation. Huck et al. (2010) and Huck and Lünser (2010) vary the amount of information available to subjects in markets with moral hazard, by varying network structure and group size. Their findings indicate partial information can be very effective at instilling trust and improving efficiency.

However, there is some reason to question how well consumer ratings will work in credence good markets, particularly when uncertainty is introduced. Mishel (1988) argues that patients seek out additional information to reduce uncertainty and the anxiety uncertainty produces. Gordon et al. (2000) find that physician disclosure of uncertainty is associated with higher patient satisfaction, but not the sole determinant. One possible mechanism is that with disclosure, when patients have a bad outcome, they give doctors the benefit of the doubt. Another mechanism is that patients, who have positive outcomes are relieved they did not have the negative one so that increases their satisfaction. In contrast, without disclosure patients are less aware of the potential negative outcome so do not have the same sense of relief. However, all this says is patients would rather know about the uncertainty, when it is present. Presumably, they would also avoid uncertainty when possible. John et al. (2019) find that people tend to "shoot the messenger" and blame (dislike) the person who delivers news of a negative event even when the person bears no responsibility for the negative outcome. They argue it is an attempt to integrate the new event with existing belief systems, and is particularly strong when the event is a surprise or somehow contradicts existing beliefs. Filippas et al. (2019) show that customer ratings have grown more positive over time, and argue that while there have been some improvements in quality, the inflation has been driven by an increased cost (feeling bad) of giving negative feedback.

The credence goods experiment literature has almost entirely relied on experiment designs where the expert is certain about the value of their credence goods to buyers. Kerschbamer and Sutter (2017, p. 20) conclude, "[a]nother very important question – according to our opinion – concerns the effects of uncertainty in the expert's diagnosis. The laboratory experiments reviewed in this paper were all characterized by the fact that expert sellers could be expected to diagnose the buyer's needs with certainty. This is obviously a harsh assumption that is violated to different degrees in most naturally occurring credence goods markets. The most prominent example for this claim is most likely the health care market where the diagnosis of a patient's needs is very often afflicted by fairly large degrees of uncertainty." We also identify an additional source of uncertainty common in credence goods, service uncertainty. Even if a doctor accurately diagnoses a patient and prescribes a treatment, the treatment may not work for that particular patient. Similarly, a financial adviser could make investment recommendations that do not deliver the expected returns because of an unforeseeable market shock.

Our experimental design tests how these two types of uncertainty impact credence goods markets. Our experimental environment focuses on *overprovision*. There is no way our experts can overcharge. While, we allow for *underprovision*, it decreases expert earnings so do not expect it. Our environment, like Plott and Wilde (1982), includes a market entry decision and allows buyers to disregard expert advice. While "commitment" to following advice has been identified as important to market efficiency, we opted for an environment in which we had both these measures of consumer trust.

2.1 Hypotheses

Based on findings in the literature we present the following hypotheses regarding how *reputation*, *diagnostic uncertainty*, and *service uncertainty*, will impact rates of overprovision, consultation, advice following and efficiency. We also hypothesize about how uncertainty will impact ratings. These hypotheses are preregistered at `https://aspredicted.org/blind.php?x=tp4az6`.

In the **Reputation Treatment**, buyers rate their experience with an expert and that rating is seen by the next buyer matched with that expert. Thus a buyer who has been the victim of overprovision (dishonest advice) can rate their experience with an expert as unsatisfactory, which will make it less likely that the next buyer assigned to the expert will consult the expert and thereby negatively impact the expert's future earnings. Any expert considering overprovision will be deterred, because detection is certain and punishment has high potential costs. In Appendix A.1, we show that if every expert whose dishonesty is discovered is given a negative rating, and it ensures that they are never consulted again (therefore all never sell the high-need service), overprovision has lower expected payout than honesty. Buyers anticipate increased honesty and are more likely to enter markets (in which they can rate experts and see experts' ratings), so are more likely to follow expert advice.[4] Decreased overprovision and increased market entry and advice following will increase buyer earnings.

Hypothesis 1: The Reputation Treatment (relative to no reputation) will: a) decrease rates of overprovision, b) increase rates of consultation, c) increase the rates buyers select the high-need service when advised to, and d) increase buyers' welfare (earnings).

In the **Diagnostic Uncertainty Treatment**, the signals the experts receive about the buyers' type are noisy. Thus, if at the expert's recommendation, a buyer buys the high-need service, but then does not get the high benefit, the buyer does not know if the expert gave dishonest advice or honest advice on a bad signal. While the bad outcome is obvious, it is not clear that the expert is to blame, so at least some buyers will not punish the expert or will punish the expert less harshly, perhaps using the "neutral" rather than "unsatisfied" rating. Any leniency in rating dishonest experts will decrease the cost for giving dishonest advice and will lead to more dishonesty. Buyers anticipating the increased dishonesty will consult less and be less likely to follow advice. The advice itself has less value to the buyer even assuming the advice is honest; expected earning of following honest advice based on a noisy signal are lower than the earnings for following advice based on a pure signal.[5] This should also decrease rates of consulting and following advice. In Appendix A.1, we show that overprovision has a higher expected profit than honesty, if buyers adjust their reactions to ratings for the noise.

Hypothesis 2: The Diagnostic Uncertainty Treatment (relative to certainty) will: a) increase rates of overprovision, b) decrease rates of consultation, c) decrease the rates buyers select the high-need service when advised to, d) reduce the frequency that experts who overprovide receive "unsatisified" ratings, e) decrease the likelihood of consultation or advice following conditional on rating seen and f) decrease buyers' welfare (earnings).

In the **Service Uncertainty Treatment**, both types of service are stochastic and might fail, resulting in no benefit. Thus if an expert overprovides and the service fails, the buyer does not know that there was overprovision. If there is no evidence of overprovision, we do not expect "unsatisfactory" ratings. This chance of not getting caught, and not being punished decreases the expected penalty, so will increase dishonesty. Buyers anticipate the increased dishonesty and also realize the ratings are less informative so are more likely to avoid the market, and less likely to follow expert advice. In Appendix A.1, we show that the

[4] Buyers' expected final payment is 23 if experts are honest, 19 is they are dishonest (20 if they do not consult), and a linear combination of 23 and 19 if there is a mix.

[5] Buyers' expected final payment is 22 if experts are honest, 20 is they are dishonest, and a linear combination if there is a mix.

decreased likelihood of being caught, makes expected value of overprovision higher than that of honesty if buyers purchase the high-need service when an unrated expert recommends it. Increased overprovision and decreased trust will both diminish buyer welfare.

Hypothesis 3: The Service Uncertainty Treatment (relative to certainty) will: a) increase rates of overprovision, b) decrease rates of consultation, c) decrease the rates buyers selected the high-need service when advised to, d) reduce the frequency that experts who overprovide receive "unsatisified" ratings, e) decrease the likelihood of consultation or advice following conditional on rating seen and f) decrease buyers' welfare (earnings).

3 Design

Our experiment introduces two sources of uncertainty into a typical credence good experiment. Subjects are assigned to one of two roles, either "Buyer" or "Expert". Buyers are assigned a need type {high, low}, which is not revealed to them, but can seek an expert's advice about which service would best serve them. There are two types of services, a less expensive service which fulfills only the low need, and a more expensive service which fulfills both needs; net of cost, low-need buyers benefit more from the former while high-need buyers benefit more from the latter. However, the expert earns more from the more expensive service. In our game, after receiving the low cost service the buyers are unable to determine if they would have gotten greater benefit from the high-need service.

We have two uncertainty treatments. The first is Diagnostic Uncertainty; experts receive a noisy signal of buyer type. Participants were told the expert administered a test which was 100% accurate in the Certainty (baseline) Treatment and only 75% accurate in the Diagnostic Uncertainty Treatment. Experts might make an 'honest' mistake; they give the appropriate advice based on an inaccurate signal. The other is Service Uncertainty; the DIY and expert services only work $66.\overline{6}\%$ of the time. (They work 100% of the time in baseline.) We also vary if there is expert reputation. The design is 2x2x2, though the cells with both types of uncertainty were not run.

The subjects were recruited by posting the study to Prolific (`www.prolific.co`), which maintains a database of people who have volunteered to participate in online research studies. We restricted recruitment to volunteers residing in the US, and only allowed volunteers to participate in the study once. Prolific notified potential participants who could read a brief description of the study and then opt to proceed to the study's web-based software. Table 1 reports participant characteristics. Our software was custom built in PHP. Subjects were assigned a treatment and a role of either buyer or expert, both of which they maintained for the duration of the experiment. All participants received instructions (based on feedback in a classroom pilot we included an option to watch a short, narrated, video version (see Appendix A.2) of the instructions). Participants took a quiz (see appendix A.3) to test comprehension before participating in the experiment. They had to complete the quiz without error to advance to the study. If a participant answered a question(s) incorrectly, they were given an indication of which question(s) was incorrect and given unlimited chances to provide a correct answer(s). Participants who found the quiz too tedious could choose not to participate and perhaps return to Prolific to find another paid study.

Table 1: Participant Characteristics

Age Range	18-25	26-45	46-64	65+			
Obs.	241	311	65	7			
Gender	Female	Male	Other				
Obs.	357	250	17				
Education	Assoc.	Bachelor	Some College	Doctor	Some HS	HS-grad	Master
Obs.	53	206	177	23	7	75	83

Figure 1: Payouts tables for the Certainty and Diagnostic Uncertainty Treatments.

The table below summarizes your earnings and the Buyers' costs and benefits based on Buyer need type and decisions.

Buyer starts with:		10 ECU		
Buyer step 1:		No Consultation	Consultation	
Buyer step 2:		DIY	DIY	Buy Service
Buyer cost:		-0 ECU	-1 ECU	-9 ECU
Buyer benefits	Low Need Buyer	+10 ECU	+10 ECU	+10 ECU
	High Need Buyer	+10 ECU	+10 ECU	+26 ECU

The table below shows your **final payments** based on Buyer's need and decisions.

		No Consultation	Consultation	
		DIY	DIY	Buy Service
	Low Need Buyer	20 ECU	19 ECU	11 ECU
	High Need Buyer	20 ECU	19 ECU	27 ECU
	Expert	0 ECU	1 ECU	3 ECU

Experts were asked to make two decisions in the experiment: 1) what advice to give to buyers whose test indicated low need; 2) what advice to give to buyers whose test indicated high need; with the knowledge that these decisions would be applied to three rounds of buyers. Strategy method was applied to these decisions, because it was straight forward, aided in online implementation, and to simplify analysis.

For each of three rounds, buyers were randomly assigned a need type, high or low with equal probability. Buyers were randomly assigned to an expert. In the first round, buyers had to consult the expert. In later rounds, buyers had the option of not consulting and using a do-it-yourself (DIY) service. If buyers consulted, they saw the advice the expert chose to give their need type. In the uncertainty treatment, there was a random draw to determine if their diagnostic test (to determine buyer need type) was accurate. After seeing the expert's advice, buyers made the choice as to whether to buy the expert's service (or to use a DIY service). Buyers learned the result of their decision and then rated the expert {☹ Unsatisfied, 😐 Neutral, ☺ Satisfied} if they consulted the expert that round and were in the reputation treatment. If buyers did not consult, they could not buy the expert's service and their only option was the cheaper DIY alternative. In the second and third rounds, buyers saw the rating that was last given (by another buyer) to their assigned expert before making any decisions. We only use three possible ratings and show only the most recent rating to simplify the analysis.

Figure 1 is a screenshot of the payout tables for the Certainty and Diagnostic Uncertainty Treatments. The upper table provides the story (starting amount, cost of service, benefit of service) to account for the payouts. The lower table only includes the resulting final payment. We conducted a pilot in one of the authors' classes that included a class post-experiment discussion. Students indicated a desire to see the information in both formats.

Figure 2 is a screenshot of the payout tables for the Service Uncertainty Treatments. As with Figure 1, the upper table communicates the story. The Service Uncertainty Treatment has two versions of the lower table one for the cases in which the service works and a second for the cases when it does not work. Payouts are increased to maintain the same expected benefit (within rounding) of the other treatments. Note that in contrast to the other treatments when the service does not work it is impossible to detect over-provision.

Figure 2: Payouts tables for the Service Uncertainty Treatments

The table below summarizes your earnings and the Buyers' costs and benefits based on Buyer need type and decisions. **Note that the service and the DIY solution only work 2 out of 3 times.** There will be a random draw to determine if it works.

Buyer starts with:		10 ECU		
Buyer step 1:		No Consultation	Consultation	
Buyer step 2:		DIY	DIY	Buy Service
Buyer cost:		-0 ECU	-1 ECU	-9 ECU
Buyer benefits if DIY or Service **WORKS**	Low Need Buyer	+15 ECU	+15 ECU	+15 ECU
	High Need Buyer	+15 ECU	+15 ECU	+39 ECU
Buyer benefits if DIY or Service **DOES NOT WORK** for	Low Need Buyer	0 ECU	0 ECU	0 ECU
	High Need Buyer	0 ECU	0 ECU	0 ECU

Final Payments if DIY or Service WORKS (2 out of 3 times)

		No Consultation	Consultation	
		DIY	DIY	Buy Service
	Low Need Buyer	25 ECU	24 ECU	16 ECU
	High Need Buyer	25 ECU	24 ECU	40 ECU
	Expert	0 ECU	1 ECU	3 ECU

Final Payments if DIY or Service DOES NOT WORK (1 out of 3 times)

		No Consultation	Consultation	
		DIY	DIY	Buy Service
	Low Need Buyer	10 ECU	9 ECU	1 ECU
	High Need Buyer	10 ECU	9 ECU	1 ECU
	Expert	0 ECU	1 ECU	3 ECU

4 Results

The experiment was run from Jan. 27th to Feb. 4th, 2021 on Prolific.co. There were 297 experts and 324 buyers. Experts earned \$5.70 and buyers \$3.78, including a \$1.30 participation fee. Subjects were allowed 20 minutes to complete the experiment. Average completion times were 9.63 minutes for experts and 10.48 minutes for buyers. Results of experts' decisions are presented in the first subsection and results of buyers' decisions are presented in the second.

4.1 Expert Decisions

Experts make two decisions: the advice given to low-need buyers and the advice given to high-need buyers. The latter is trivial (interests are aligned) and can be used as an attention check for the former. Advising high-need buyers to DIY (which benefits neither party) occurred at just under 5%, indicating good participant comprehension. In the former, interests are not aligned; experts can advise buyers to buy their service and increase their own earnings at the expense of buyers. This is commonly referred to as overprovision. Our analysis of expert decisions focuses on how overprovision is impacted by uncertainty and reputation. Figure 3 reports rates of overprovision (expert dishonesty) by treatment. Dashed lines indicate χ^2-test; p-values for the correspond tests are beside lines.[6] The χ^2-test results support Findings 1a, 2a, 3a.

Figure 3: Rates of Overprovision by Treatment

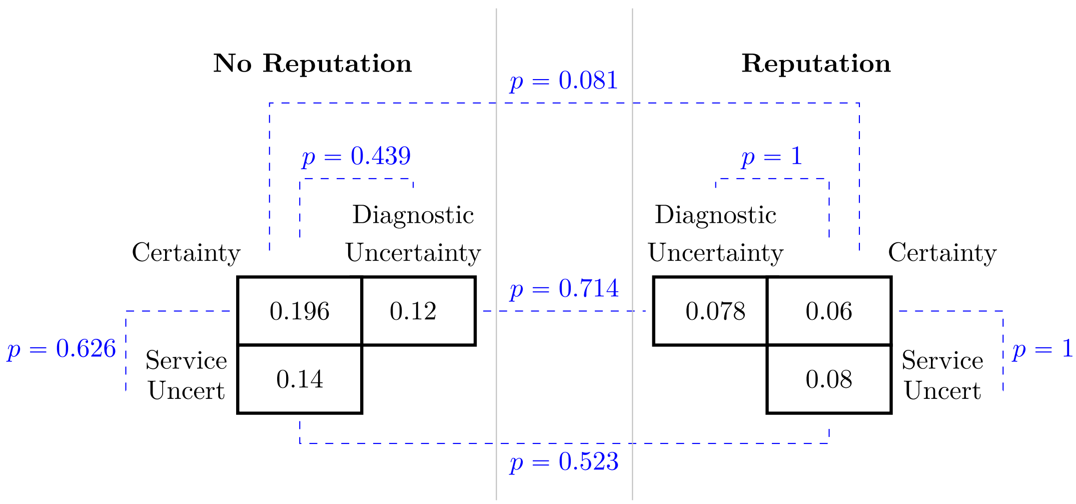


Table 2: Expert Earnings by Treatment and Overprovision, Balanced

Rep	ServUncert	DiagUncert	ECu's Over	ECU's Truth	LiePay	Over
0	0	0	1.819	1.612	0.207	0.196
0	0	1	2.065	1.719	0.347	0.120
0	1	0	2.214	1.643	0.571	0.140
1	0	0	2.205	1.588	0.618	0.060
1	0	1	1.381	1.464	-0.084	0.078
1	1	0	1.773	1.431	0.342	0.080

Table 2 reports expert earnings by treatment and overprovision versus truthfully reporting need level, fixing buyer need type at the expected 50-50 distribution. It was constructed by calculating mean earnings

[6] As none of the tests are statistically significant at conventional levels, we do not correct for multiple hypothesis testing.

for each treatment arm and buyer need level, and then taking mean of means, ensuring equal weight to each need level in every treatment. *LiePay* is the difference between the proceeding two columns. Overprovision was profitable in all but one treatment, Diagnostic Uncertainty with Reputation. With the possible exception of the certainty treatments, expected overprovision was more profitable in the treatments without reputation. Within no reputation, dishonesty is more profitable with uncertainty, product uncertainty more so. The final column reports the proportion of experts who recommended their service to low-need buyers.

Finding 1a: As hypothesized, the Reputation Treatments, in which buyers rate experts and can avoid experts with negative ratings, have lower rates of overprovision than the No Reputation Treatments. The p-value of a χ^2-test pooling all three cells is 0.045. However, the difference across individual cells is not statistically significant for $p < 5\%$

Finding 2a: Contrary to our hypothesis, differences in rates of overprovision are not statistically significant between diagnostic uncertainty (where the buyer cannot distinguish between overprovision and an inaccurate test) and certainty (where the low payout only occurs when the expert is dishonest). Rates are actually lower when there is no reputation.

Finding 3a: Contrary to our hypothesis, differences in rates of overprovision are not statistically significant between service uncertainty (where when the service does not work overprovision is undetectable) and certainty. Rates are actually lower when there is no reputation.

4.2 Buyer Decisions

Buyers make make up to three decisions each round: whether or not to buy the experts' service, and in the reputation Treatments how to rate the expert, in the second and third rounds, whether or not to consult. In the first round s must consult and making a decision about buying without any personal history or rating information. We analyze the first round buying decision first, then analyze rating decisions, and finally analyze decisions that are impacted by ratings, consultation and later round buying decisions.

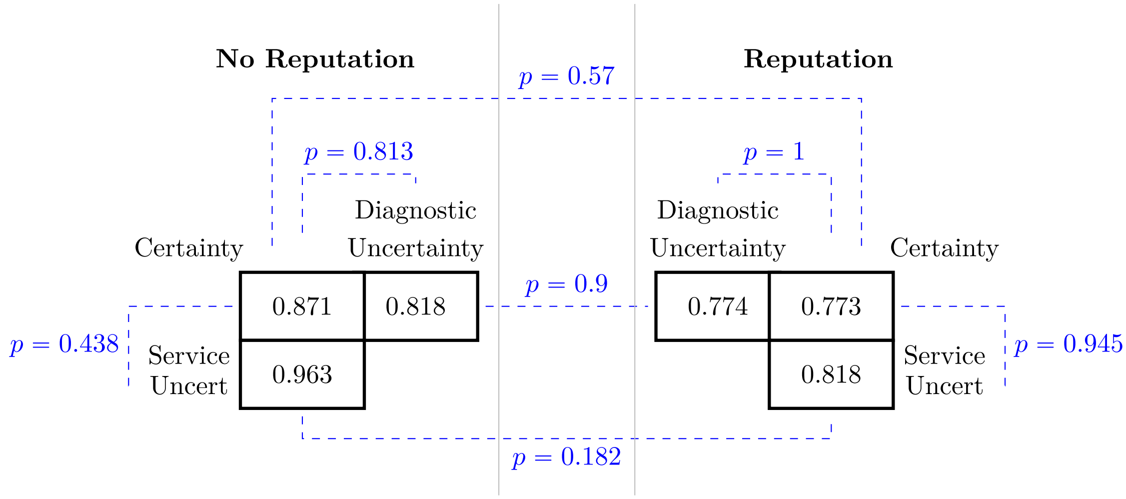
4.2.1 First round buying

Figure 4 reports rates of taking advice to "Buy" by treatment, during Round 1 when there are no ratings. Dashed lines indicate χ^2-test; p-values for the correspond tests are beside lines. The χ^2-test results support Findings 1c1, 2c1, 3c1.

Finding 1c1: Within the first round, contrary to our hypothesis, the Reputation Treatments, in which experts have greater incentives to be honest, have lower rates of accepting advice to buy than the No Reputation Treatments. The p-value of a χ^2-test pooling all three cells is 0.003. However, the difference across individual cells is not statistically significant.

Finding 2c1: Within the first round, contrary to our hypothesis, differences in rates of accepting advice to buy are not statistically significant between diagnostic uncertainty (where the incentives to be honest are weakened by fact deliberate overprovision cannot be identified) and certainty.

Figure 4: Rates of Taking Round 1 Advice to "Buy" by Treatment



Finding 3c1: Within the first round, contrary to our hypothesis, differences in rates of accepting advice to buy are not statistically significant between service uncertainty (where the incentives to be honest are weakened by the fact that when the service does not work it obscures overprovision) and certainty. However the differences are not statistically significant.

4.2.2 Ratings

The proportions of each rating {Unsatisfied, Neutral, Satisfied} given by buyers following advice for each treatment, are displayed in Figure 5. Advice was followed 79% of the time. We omit rating from buyers not following advice because, the relationship between outcome and rating given should be substantially different depending upon whether the advice was followed, perhaps being inverse. p-values from χ^2 tests of ratings across treatments are shown above bars. There are many more Unsatisfied ratings in Service Uncertainty Treatment than in the other two treatments. Differences between Certainty and Diagnostic Uncertainty are not statistically significant. Chi values and further tests are reported in Table A.4. Table A.3 reports counts. To understand what is driving the treatment differences, we examine within each treatment how ratings varied by outcome.

Figure 5: Rating by Treatment

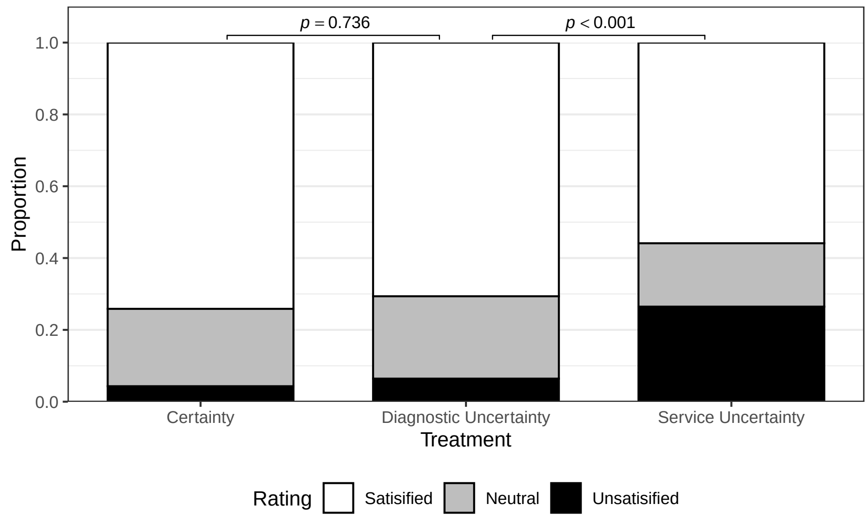


Figure 6 displays ratings given by buyers following the experts' advice in the Certainty Treatment. Each bar represents a particular level of earnings (bottom axis), and shows proportions of each of the three ratings within that level. Advice and type (if the outcome reveals it) are on top. The ratings, given the outcomes, are much as expected. The majority of buyers, who bought when it was not in their best interest, and earn 11 ECU, rate the experts, who advised them to do so, negatively. The majority of buyers with both the other outcomes rate the experts who gave them honest advice positively. p-values from χ^2 tests of ratings across outcomes are shown above bars. Chi values and the test of the final combination are reported in Table A.6. Table A.5 reports counts.

Figure 6: Rating by Outcome in Certainty Treatment

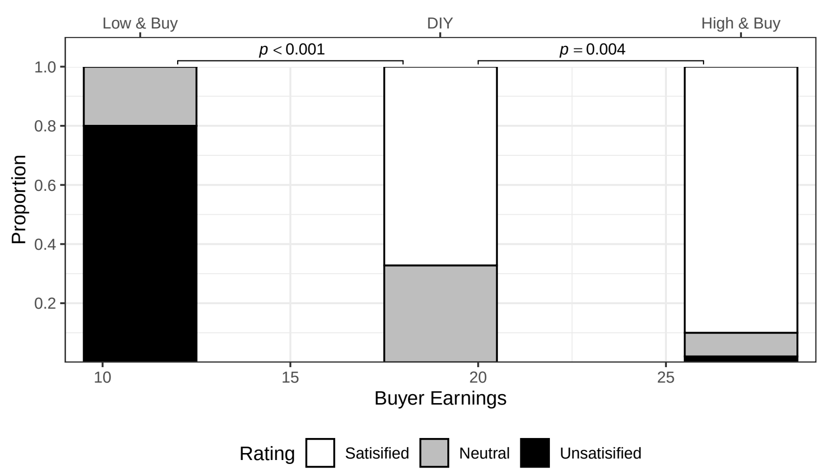


Figure 7 displays ratings given by buyers following the experts' advice in the Diagnostic Uncertainty. Each bar represents a particular level of earnings (bottom axis), and shows proportions of each of the three rating within that level. Advice and type (if the outcome reveals it) are on top. buyers who get 11 ECUs appear to give experts the benefit of doubt that the bad of the outcome is from the test being wrong rather than the expert's greed, which would explain the relative lack of Unsatisfied ratings. However, we expected Neutral ratings would be more common than Satisfied ratings in this situation. Otherwise the ratings for given outcomes follow the same pattern as in the Certainty Treatment. p-values from χ^2 tests of ratings across outcomes are shown above bars. Chi values and the test of the final combination are reported in Table A.8. Table A.7 reports counts.

Figure 7: Rating by Outcome in Diagnostic Uncertainty Treatment

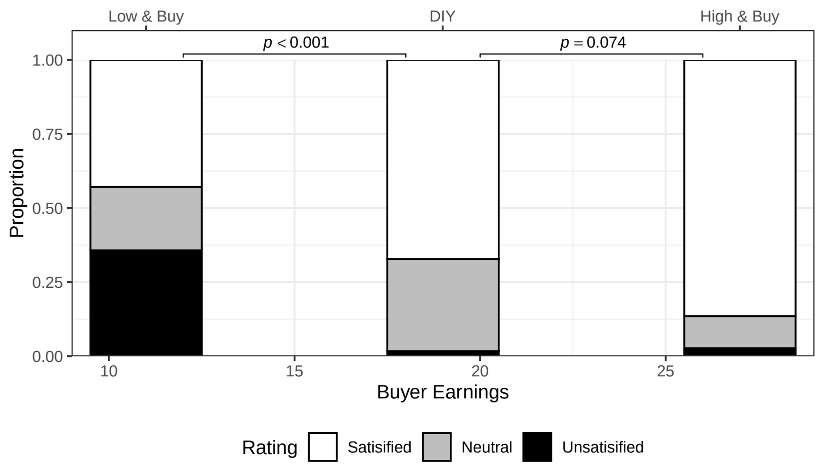


Figure 8 displays ratings given by buyers following the experts' advice in the Service Uncertainty Treatment. Each bar represents a particular level of earnings (bottom axis), and shows proportions of each of the three rating within that level. Advice, not working, and type (if the outcome reveals it) are on top. Buyers clearly punish experts with Unsatisfied ratings when the service they bought does not work, despite that whether it worked was entirely random and independent of the expert. Curiously, there is not a similar punishment when DIY does not work. we do not see evidence of punishment when low-need buyers bought the service. Chi values and further tests are reported in Table A.10. Table A.9 reports counts.

Figure 8: Rating by Outcome in Service Uncertainty Treatment

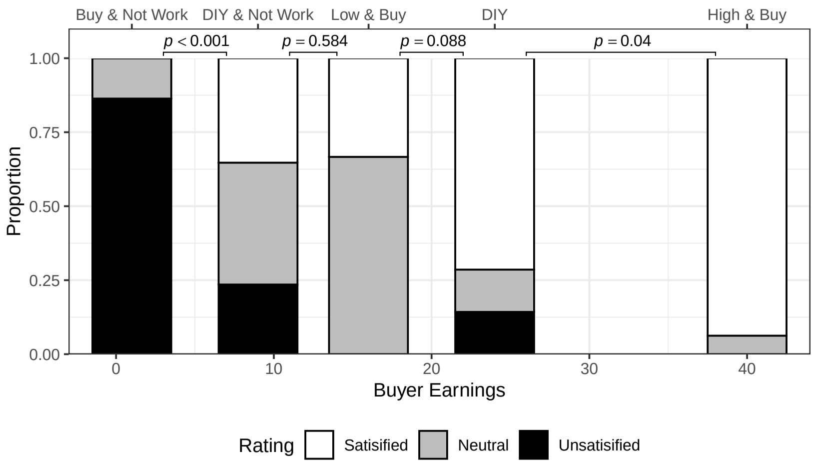


Table 3 reports estimated marginal effects from an ordered probit regression on the ratings given by buyers who follow experts' advice. Table A.11 reports the coefficient estimates and cut points. Errors are clustered on buyers. The reference cell is certainty with truthful advice (not overprovision). The first row shows, in certainty, overprovision is penalized; relative to the reference cell ☹ Unsatisfied is 71% more likely, while ☺ Satisfied is 67% less likely. Both are statistically significant. The remaining 4% is accounted for by a decrease in the 😐 Neutral ratings. Penalties are less likely (there are smaller marginal effects) for overprovision in the Diagnostic Uncertainty and Service Uncertainty Treatments, than the Certainty Treatment. They are also less severe; the statistically significant increase is for the Neutral rating, not Unsatisfied. The *Service Fails* variable is binary and takes the value 1 when the service does not work. The estimates show that buyers penalize the expert when the Service does not work, despite that this was random and independent of the expert. This confirms what was graphically depicted in Figure 8. Table A.12 reports marginal effect from a model without the *Service Fails* variable.

Table 3: Estimated Marginal Effects from an Ordered Probit Regression on Rating

	☹	😐	☺
Certainty Overprovision	0.713***	-0.0398	-0.674***
	(0.165)	(0.116)	(0.0622)
Diagnostic Uncertainty & Truthful	0.0104	0.0150	-0.0254
	(0.0229)	(0.0326)	(0.0554)
Diagnostic Uncertainty & Overprovision	0.245	0.159***	-0.405**
	(0.130)	(0.0276)	(0.139)
Service Uncertainty & Truthful	-0.00826	-0.0125	0.0207
	(0.0262)	(0.0404)	(0.0666)
Service Uncertainty & Overprovision	0.154	0.139**	-0.292*
	(0.0878)	(0.0470)	(0.131)
Service Fails	0.269***	0.316***	-0.585***
	(0.0365)	(0.0540)	(0.0751)
Observations	327	327	327

Clustered robust Std. Err. in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finding 2d: Consistent with our hypothesis, when there was diagnostic uncertainty (and it was unclear if overprovision was deliberate or the result of an inaccurate test) buyers punished overprovision less harshly, reducing the reliability of ratings.

Finding 3d: Consistent with our hypothesis, when there was service uncertainty and the high-need service did not work, buyers were less likely to give satisfied ratings. Additionally, to our surprise, buyers were less likely to punish overprovision when the service worked. Both of these reduced the reliability of ratings.

4.2.3 Consulting

Table 4 reports estimated marginal effects from a probit regression of likelihood of consulting. There are no statistically significant differences between consultation rates due to uncertainty or reputation. Buyers in Certainty & Reputation are 3% less likely to consult relative to the omitted cell (Certainty & No Reputation). Buyers in Diagnostic Uncertainty & No Reputation are 2% less likely to consult relative to Diagnostic Uncertainty & Reputation. Buyers in Product Uncertainty & No Reputation are 10% less likely to consult relative to Product Uncertainty & Reputation. This difference is not statistically significant. The result is driven by negative ratings after BUY/DIY doesn't work (see below).

Table 4: Estimated Marginal Effects from a Probit Regression of Likelihood to Consult by Treatment

	(1)	
Treatment:		
Certainty & Reputation	-0.0328	(0.0725)
Diagnostic Uncertainty & Reputation	-0.0789	(0.0696)
Diagnostic Uncertainty & No Reputation	-0.0557	(0.0679)
Service Uncertainty & No Reputation	-0.0471	(0.0668)
Service Uncertainty & Reputation	-0.150*	(0.0642)
Observations	649	
Buyers	325	
Log Pseudolikelihood	-402.2	

Clustered robust Std. Err. in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5 reports estimated marginal effects of how ratings and treatment impact likelihood to consult an expert. Ratings seen is the expert's reputation. Last rating given is the rating that the buyer (that is making the consultation decision) gave their expert last round or in round 1, if they did not consult in Round 2. We include it to test how much the buyers' experiences impact their decisions. In Model 1, ratings seen have expected impact on consultation. Buyers are 35% more likely to consult an expert with the rating of 😐 Neutral and 61% more likely to consult an expert with the rating of 🙂 Satisfied, relative to an expert ☹ Unsatisfied. These effects vary slightly across specifications. Model 2 shows that buyers who were Unsatisfied with their previous experience are 16% less likely to consult. In mode1s 3 and 4, like Table 4 Diagnostic Uncertainty & Reputation is not different than Certainty & Reputation. Here Product Uncertainty & Reputation is also not different, indicating it was due to the decrease in Satisfied ratings. The results in Tables 5 support the following findings 1b, 2b and 2c.

Finding 1b: Contrary to our hypothesis, consultation rates do not increase when ratings are displayed and actually decrease in the service uncertainty treatment, due to negative ratings when the services do not work.

Table 5: Estimated Marginal Effects of Likelihood to Consult

	(1)	(2)	(3)	(4)
Rating Seen 😐	0.345***	0.318***	0.363***	0.344***
	(0.0726)	(0.0765)	(0.0730)	(0.0766)
Rating Seen ☺	0.605***	0.568***	0.623***	0.594***
	(0.0621)	(0.0672)	(0.0626)	(0.0674)
Last Rating Given 😐		0.0832		0.113
		(0.0818)		(0.0864)
Last Rating Given ☺		0.159*		0.187*
		(0.0738)		(0.0791)
Diagnostic Uncertainty			-0.00375	-0.00354
			(0.0661)	(0.0646)
Service Uncertainty			0.0519	0.0788
			(0.0656)	(0.0659)
Observations	325	325	325	325
Buyers	163	163	163	163
Log Pseudolikelihood	-171.6	-169.1	-171.1	-168.1

Clustered robust Std. Err. in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finding 2b: Contrary to our hypothesis, the difference in consultation rates between diagnostic uncertainty (in which inaccurate test results and the increased likelihood of dishonest experts depreciate the value of advice) and certainty are not statistically significant.

Finding 3b: Contrary to our hypothesis, consultation rates do not decrease when there service uncertainty (relative to service certainty). Contrary to predictions, the difference in consultation rates between service uncertainty (in which the increased likelihood of dishonest experts depreciate the value of advice) and certainty are not statistically significant.

Table 6 reports marginal effects of a probit regression testing if there are differences across treatments in how rating seen impacts likelihood of consulting. The reference case is an Unsatisfied rating in the Certainty Treatment. While relative to that case, buyers who see an Unsatisfied rating in the other treatments are 45% and 28% less likely to consult, none of the differences are statistically significant. The differences across treatments for the other two ratings are much smaller $\leqslant 7\%$.

Table 6: Estimated Marginal Effects of Likelihood to Consult by Treatment and Rating Seen

	Consult	
Diagnostic Uncertainty ☹	-0.450	(0.239)
Service Uncertainty ☹	-0.279	(0.238)
Certainty 😐	0.0137	(0.238)
Diagnostic Uncertainty 😐	0.0619	(0.244)
Service Uncertainty 😐	0.0795	(0.249)
Certainty ☺	0.284	(0.241)
Diagnostic Uncertainty ☺	0.304	(0.235)
Service Uncertainty ☺	0.354	(0.234)
Observations	325	
Buyers	163	
Log Pseudolikelihood	-168.8	

Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.2.4 Buying

Table 7 reports marginal effects from probit regressions for rounds when the expert gave the advice to buy the high-need service. The dependent variable takes the value 1 if the buyer follows the advice. Model 1 only includes Period 1 (no clustering) and shows there are no statistically significant differences across treatments.[7] The results are similar to the results of t-tests in Figure 4. Model 2 includes all periods; the estimates are consistent with previous model. Model 3, adds ratings; while there is still no treatment effect, relative to Unsatisfied, Neutral and Satisfied are statistically significantly more likely to buy the high-need service. These support findings 1c2, 2c2 and 3c2:

[7] The estimates for *Diagnostic Uncertainty No Rep* and *Service Uncertainty Rep* are identical because both cells have an identical distribution of outcomes.

Table 7: Estimated Marginal Effects from Probit Regressions of Likelihood to Buy

	(1)	(2)	(3)
Treatments:			
Certainty Rep	-0.0982	0.0309	
	(0.108)	(0.0801)	
Diagnostic Uncertainty Rep	-0.0968	-0.0977	-0.113
	(0.0963)	(0.0739)	(0.0897)
Diagnostic Uncertainty No Rep	-0.0528	0.0125	
	(0.0902)	(0.0730)	
Service Uncertainty No Rep	0.0920	0.0787	
	(0.0703)	(0.0635)	
Service Uncertainty Rep	-0.0528	-0.103	-0.0556
	(0.0902)	(0.0695)	(0.0968)
Rating Seen 😐			0.265*
			(0.104)
Rating Seen 🙂			0.532***
			(0.0955)
Observations	177	541	168
Buyers		268	108
Log Likelihood	-75.84		
Log Pseudolikelihood		-325.1	-96.87

Clustered robust Std. Err. in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finding 1c2: Contrary to our hypothesis, the difference in rates of following advice to "Buy" between the Reputation Treatments (where ratings increase the cost of expert dishonesty) and the No Reputation Treatments are not statistically significant.

Finding 2c2: Contrary to our hypothesis, the difference in rates of following advice to "Buy" when there is diagnostic uncertainty (and the advice may be based on an inaccurate test) are not statistically significant from when there is certainty.

Finding 3c2: Contrary to our hypothesis, controlling for the expert's rating, the difference in rates of following advice to "Buy" when there is service uncertainty (and the service not working may obscure dishonesty) are not statistically significant from when there is certainty. experts in the service uncertainty treatment are more likely to have a negative rating.

Table 8 reports marginal effects for a probit regression testing if there are differences across treatments in how rating seen impacts likelihood of buying the high-need service. The reference case is an Unsatisfied rating in the Certainty Treatment. While relative to that case, buyers who see an Unsatisfied rating in the other treatments are 50% and 40% less likely to consult, none of the differences are statistically significant.

Finding 2e: Contrary to our hypothesis, when there was diagnostic uncertainty and the reliability of ratings was reduced, the change in the likelihood of consulting or buying for the rating seen was not statistically significant.

Table 8: Estimated Marginal Effects of Likelihood to Buy by Treatment and Rating Seen

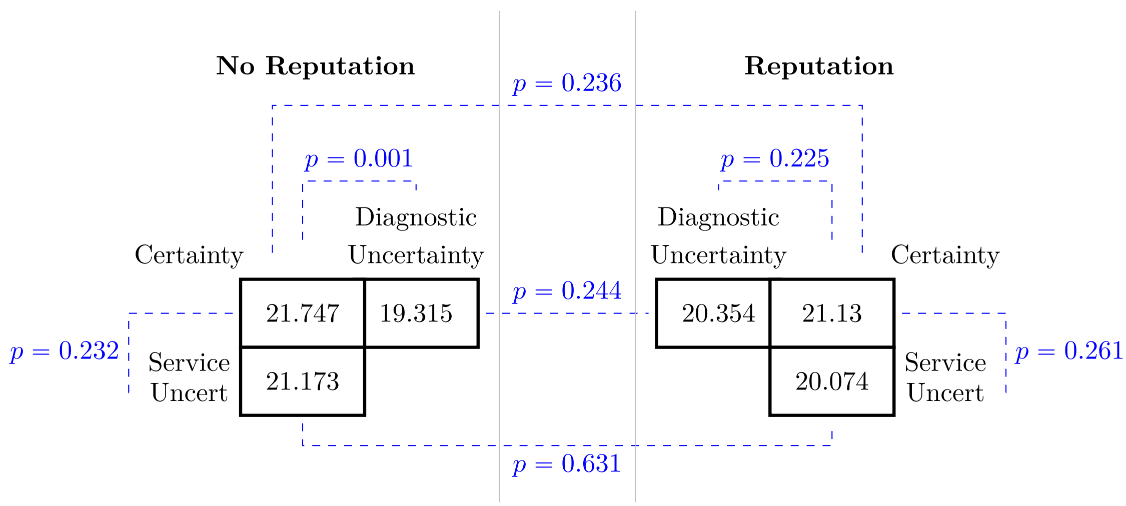
	Buy	
Diagnostic Uncertainty ☹	-0.500	(0.279)
Service Uncertainty ☹	-0.398	(0.268)
Certainty 😐	0.0961	(0.288)
Diagnostic Uncertainty 😐	-0.287	(0.272)
Service Uncertainty 😐	-0.0174	(0.280)
Certainty ☺	0.150	(0.273)
Diagnostic Uncertainty ☺	0.244	(0.262)
Service Uncertainty ☺	0.136	(0.270)
Observations	168	
Buyers	108	
LogPseudolikelihood	-92.79	

Clustered robust Std. Err. in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finding 3e: Contrary to our hypothesis, when there was service uncertainty and the reliability of ratings was reduced, the change in the likelihood of consulting or buying given the the buyer rating seen was not statistically significant.

Figure 9 displays mean buyer ECU by treatment. Dashed lines show Wilcoxon signed-rank tests and their p-values. For comparison, if experts were completely trustworthy (there was no overprovision) and buyers were completely trusting (always consulted and always followed advice), buyers' ECUs would be 23; half the buyers would be low type and earn 19 ECUs and the other half would be high and earn 27. In service uncertainty, a sixth earns 9 another sixth earns 1, a third earns 24 and the final third earns 40.

Figure 9: Mean Buyer ECU by Treatment



Finding 1d: Contrary to our hypothesis, differences in buyer earnings between the Reputation Treatments (which theoretically increases the cost of expert dishonesty and increases buyer trust) and the No Reputation Treatments are not statistically significant.

Finding 2f: Consistent with our hypothesis, buyers' ECUs are lower (statistically significant) with diagnostic uncertainty when there is no reputation. They are also lower when there is reputation but the difference is not statistically significant.

Finding 3f: Contrary to our hypothesis, buyers' ECUs are not lower (statistically significant) with service uncertainty.

Figure 10: Mean Expert ECUs by Treatment

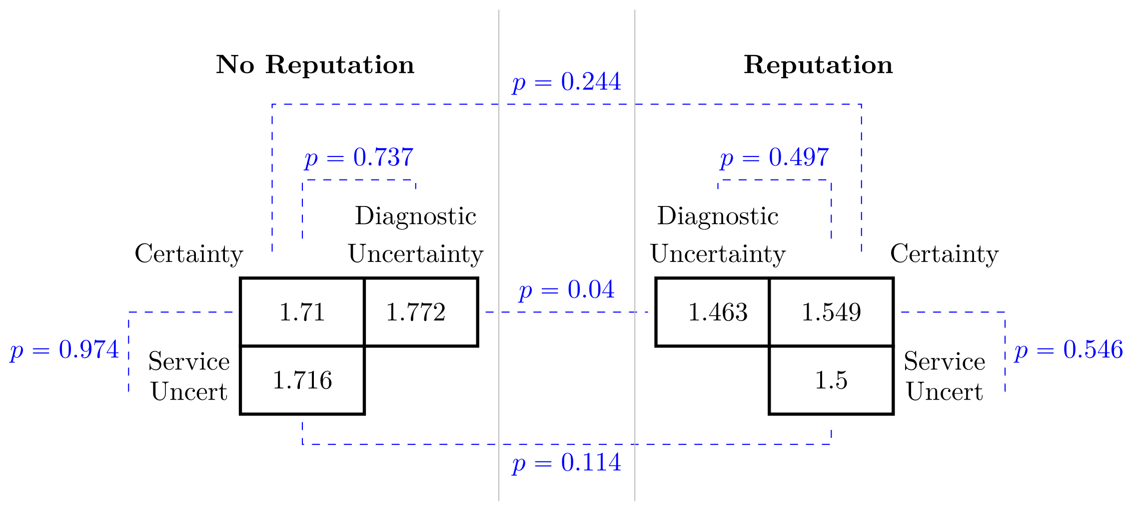


Figure 10 displays mean expert earnings in ECUs per round by treatment. Dashed lines show Wilcoxon signed-rank tests and their p-values. While none of the cross cell tests are statistically significant, the overall (all cells pooled) test of reputation is (p-value=0.005). For comparison, if experts were completely trustworthy and buyers were completely trusting, experts' ECUs would be 2; half the buyers would be low type and consult earning the expert 1 ECU and the other half would be high and buy earning the expert 3. If experts always overprovided and buyers were completely naive, and always consulted and always bought, earnings would be 3. If buyers completely exited the market and did not buy and did not consult after the first round, ECUs would be 0.33.

5 Discussion

In summary, despite the decreased likelihood of being caught, we find no evidence that either diagnostic or service uncertainty results in increased dishonesty by experts, nor any decreased trust from buyers. Within reputation, there are (not statistically significant) increases in over-provision for both types of uncertainty. Within no reputation, there were decreases. It is possible that experts felt more sympathy for buyers when uncertainty and a lack of ratings complicated the choice and made it more difficult. It may also be that ratings provide a moral licensing, i.e. an expert might think it's unfair to use uncertainty to exploit the helpless buyer; however, if the buyer can make and view rating they are no longer helpless, making it fair game.

The slightly lower (not statistically significant) consultation rate with diagnostic uncertainty, that we find throughout the analyses, is consistent with the lower value of the information. In the first round buy decision, buyers followed advice at higher rates in the no reputation cells. This is surprising because, though the buyers in the reputation cells did not see a rating to inform their decision, they had the assurance that they could punish the expert if following the advice led to a negative outcome (as the experts are aware). Even in later rounds, advice to buy was followed at lower rates in the reputation treatments. This rules out the possibility, that buyers were cautious in the first round, but more trusting after ratings were available.

Ratings were noisy, particularly when there was uncertainty, so limited the value of reputation. With diagnostic uncertainty, when the service failed to work, buyers seem to give the expert the “benefit of the doubt”; satisfied ratings were more common than neutral ratings in these situations. This finding is consistent with Filippas et al.’s (2019) assertion that customers eschew unfavorable ratings to avoid feeling bad for giving negative feedback. It may be possible that a stylization we made contributed to this reluctance. In order to simplify the design and analysis, we chose to only display the most recent rating the expert had received. We could have chosen to average all the ratings, which may decrease the potential for guilty feelings for not giving an expert top ratings. With service uncertainty, reputation was detrimental; when service failed, experts, including those who were honest, were given negative ratings, which then deterred buyers from consulting or receiving services from these experts. This finding is consistent with John et al.’s (2019) finding that participants “shoot the messenger” and punish the bearers of bad news for reporting bad outcomes that were determined independent of the messenger. However, when the buyers chose DIY and it failed, they did not blame the expert. This may be because the expert did not earn any pay for the choice of DIY, or that the framing gives the buyer agency over the outcome. However, we note that the expert does not actually have a messenger role. Experts made decision hours before buyers via strategy method; software simply implement the decisions and reported the earnings. Buyers were not directly told whether either service did not work, though could infer that from their payout and the payout table(s), which were visible when they rated the expert.

Despite that ratings in the uncertainty treatments were noisier and had less information, they seemed to have no less impact on the next buyer’s consult and buy decisions. It is possible that greater uncertainty may magnify the buyer’s sense of need for information and solutions (Mishel, 1988), driving expert consultation, and trust in expert advice, which offset any discounting. However, our experiment was not designed to detect or isolate the effect of the potentially competing mechanisms.

Given that there were few statistically significant findings among the decision variables, the lack of statistically significant differences in buyer earnings is not surprising. The difference within no reputation between certainty and diagnostic uncertainty is puzzling because the findings thus far indicate no mechanism for the difference. The most likely cause seems to be that the realization of the random draw for the test result varies from the expected mean; it is 42% not 33%. It may be due to the mechanics of the design; we didn’t change payout to compensate, for bad signals leading to bad decisions and decreased payouts. experts earn more without reputation because they are less honest and over-provide more when there is no reputation.

There is understandable skepticism of results from online platforms. Final payments for our study were less than typical lab experiments; total pay was below the show-up fee for our institution’s lab. However, hourly pay was on average \$59/hour for experts and \$36/hour for buyers, so above our lab’s standard and well above Prolific’s standard. We are confident this incentive level and our comprehension test ensures quality responses. Our design also includes two measures of participant mistakes. Experts can under-provide service–recommend DIY to the high-need type. However, this decrease both their own earnings and those of buyer. The rate of under-provision is less than 5%, so indicates the vast majority of participants were sufficiently attentive. Similarly, buyers can reject the advice to DIY. Essentially, this is believing the expert was trying to under-provision. Recognizing this as a mistake is $k+1$ order from under-providing and occurs a higher rate but is still bellow 8%. Our sample size is moderate so might explain the lack of evidence. In the no reputation conditions, there were lower rates of overprovision, so our data suggest that an increased sample size is more likely to result in statistical significance of an effect in the opposite direction than predicted. However, within reputation, it is possible that a larger sample size might result in a statistically significant difference.

Our results are consistent with Balafoutas et al. (2020), who find that diagnostic uncertainty reduces efficiency. Our results are also similar to Jin et al. (2021) who find senders do not fully exploit the ability to withhold information and receivers are "insufficiently skeptical about undisclosed information" (p. 143). We, like Deck and Tracy (2020), find that buyers may rely too heavily on ratings when deciding how much trust to place in experts, despite the limitations of the ratings and failing to utilize other sources of information. While tentative, our results imply that social norms and concern for a positive self-image may discipline credence goods experts even when uncertainty obfuscates deception. Our findings also suggest that while ratings systems are imperfect, and may not flag every bad actor, or properly weigh information–their mere existence might provide sufficient deterrence of misdeeds.

References

Akerlof, George A., "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, 1970, *84*, 488–500. 3.

Arrow, Kenneth J., "Uncertainty and the welfare economics of medical care," *The American economic review*, 1963, *53* (5), 941–973.

Balafoutas, L., A. Beck, R. Kerschbamer, and M. Sutter, "What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods," *The Review of Economic Studies*, July 2013, *80* (3), 876–891.

Balafoutas, Loukas, Helena Fornwagner, Rudolf Kerschbamer, Matthias Sutter, and Maryna Tverdostup, "Diagnostic Uncertainty and Insurance Coverage in Credence Goods Markets," SSRN Scholarly Paper ID 3726066, Social Science Research Network, Rochester, NY November 2020.

_ , Rudolf Kerschbamer, and Matthias Sutter, "Second-Degree Moral Hazard in a Real-World Credence Goods Market," *The Economic Journal*, February 2017, *127* (599), 1–18.

Beck, Adrian, Rudolf Kerschbamer, Jianying Qiu, and Matthias Sutter, "Car mechanics in the lab—Investigating the behavior of real experts on experimental markets for credence goods," *Journal of Economic Behavior & Organization*, December 2014, *108*, 166–173.

Bejarano, Hernán, Ellen P. Green, and Stephen Rassenti, "Payment scheme self-selection in the credence goods market: An experimental study," *Journal of Economic Behavior & Organization*, October 2017, *142*, 396–403.

Bester, Helmut and Matthias Dahm, "Credence Goods, Costly Diagnosis and Subjective Evaluation," *The Economic Journal*, June 2018, *128* (611), 1367–1394.

Bohnet, Iris and Steffen Huck, "Repetition and Reputation: Implications for Trust and Trustworthiness When Institutions Change," *American Economic Review*, April 2004, *94* (2), 362–366.

Cabral, Luís and Ali Hortaçsu, "The Dynamics of Seller Reputation: Evidence from Ebay*," *The Journal of Industrial Economics*, 2010, *58* (1), 54–78. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6451.2010.00405.x.

Crawford, Vincent P. and Joel Sobel, "Strategic Information Transmission," *Econometrica*, 1982, *50* (6), 1431–1451. Publisher: [Wiley, Econometric Society].

Darby, Michael R. and Edi Karni, "Free competition and the optimal amount of fraud," *The Journal of law and economics*, 1973, *16* (1), 67–88.

Deck, Cary and J. Tracy, "Keeping a Clean Reputation: More Evidence on the Perverse Effects of Disclosure," ESI Working Paper 20-21, Chapaman University November 2020.

Dulleck, Uwe and Rudolf Kerschbamer, "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods," *Journal of Economic Literature*, February 2006, *44* (1), 5–42.

_ , _ , and Matthias Sutter, "The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition," *American Economic Review*, April 2011, *101* (2), 526–555.

Filippas, Apostolos, John J. Horton, and Joseph M. Golden, "Reputation Inflation," Technical Report w25857, National Bureau of Economic Research May 2019.

Gillies, Anthony S. and Mary L. Rigdon, "Plausible Deniability and Cooperation in Trust Games," *Review of Behavioral Economics*, 2019, *6* (2), 95–118. Publisher: Now Publishers, Inc.

Gneezy, Uri, "Deception: The Role of Consequences," *American Economic Review*, March 2005, *95* (1), 384–394.

Gordon, Geoffrey H, Sandra K Joos, and Jennifer Byrne, "Physician expressions of uncertainty during patient encounters," *Patient Education and Counseling*, April 2000, *40* (1), 59–65.

Gottschalk, Felix, Wanda Mimra, and Christian Waibel, "Health Services as Credence Goods: a Field Experiment," *The Economic Journal*, July 2020, *130* (629), 1346–1383.

Huck, Steffen and Gabriele K. Lünser, "Group reputations: An experimental foray," *Journal of Economic Behavior & Organization*, February 2010, *73* (2), 153–157.

_ , _ , and Jean-Robert Tyran, "Consumer networks and firm reputation: A first experimental investigation," *Economics Letters*, August 2010, *108* (2), 242–244.

Jin, Ginger Zhe, Michael Luca, and Daniel Martin, "Is No News (Perceived As) Bad News? An Experimental Investigation of Information Disclosure," *American Economic Journal: Microeconomics*, May 2021, *13* (2), 141–173.

John, Leslie, Hayley Blunden, and Heidi Liu, "Shooting the Messenger," *Journal of Experimental Psychology: General*, April 2019, *148* (4), 644–666.

Kerschbamer, Rudolf and Matthias Sutter, "The Economics of Credence Goods – a Survey of Recent Lab and Field Experiments*," *CESifo Economic Studies*, March 2017, *63* (1), 1–23.

_ , _ , and Uwe Dulleck, "How Social Preferences Shape Incentives in (Experimental) Markets for Credence Goods," *The Economic Journal*, March 2017, *127* (600), 393–416.

Liu, Fang, Alexander Rasch, Marco Schwarz, and Christian Waibel, "The Role of Diagnostic Ability in Markets for Expert Services," *SSRN Electronic Journal*, 2019.

Luca, Michael, "Reviews, reputation, and revenue: The case of Yelp.com," Harvard Business School NOM Unit Working Paper 12-016 2016.

Mishel, Merle H., "Uncertainty in Illness," *Image: the Journal of Nursing Scholarship*, 1988, *20* (4), 225–232. _eprint: https://sigmapubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1547-5069.1988.tb00082.x.

Pitchik, Carolyn and Andrew Schotter, "Honesty in a Model of Strategic Information Transmission," *The American Economic Review*, 1987, *77* (5), 1032–1036.

Plott, Charles R and Louis L Wilde, "Toward A Theory of Professional Diagnosis and Service: Consumer Behavior," Technical Report 352, Califomia Institute of Technology, Pasadena,CA 1980.

_ and _ , "Professional Diagnosis vs. Self-Diagnosis: An Experimental Examination of Some Special Features of Markets with Uncertainty," in Vernon L. Smith, ed., *Research in Experimental Economics, Volum*, Vol. 2, JAi Press Inc, 1982, pp. 63–112.

Reimers, Imke and Joel Waldfogel, "Digitization and Pre-purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings," *American Economic Review*, June 2021, *111* (6), 1944–1971.

Rowell, David and Luke B. Connelly, "A history of the term "moral hazard"," *Journal of Risk and Insurance*, 2012, *79* (4), 1051–1075.

Rubin, Jared and Roman Sheremeta, "Principal–Agent Settings with Random Shocks," *Management Science*, April 2016, *62* (4), 985–999.

Schneider, Henry S., "Agency Problems and Reputation in Expert Services: Evidence from Auto Repair," *The Journal of Industrial Economics*, September 2012, *60* (3), 406–433.

Sánchez-Pagés, Santiago and Marc Vorsatz, "An experimental study of truth-telling in a sender–receiver game," *Games and Economic Behavior*, October 2007, *61* (1), 86–112.

Wolinsky, Asher, "Competition in a Market for Informed Experts' Services," *The RAND Journal of Economics*, 1993, *24* (3), 380–398. Publisher: [RAND Corporation, Wiley].

Appendices

A.1 Profitability of Overprovision

C is profit if the buyer consults. B is profit if the buyer buys. λ and μ are the expert's respective beliefs about probability the buyer will consult, and buy the expert's service if advised. The subscript 0 denotes when there is no reputation, i.e. round 1, $\ominus$ denotes events after the expert has been caught in lying to a buyer, and $\oplus$ is when the expert has not been caught in lie (or is honest).

A.1.1 Certainty

In the certainty cells, the experts receive a precise signal of buyer type, and services always work without fail, so expert dishonesty is the only reason a buyer following expert advice would not get the best outcome possible given their type.

Lemma 1 Given $C = 1$ and $B = 3$, the net profit from overprovision is,

$$\mathbb{E}[\pi(Lie) - \pi(Truth)] = \mathbb{E}[\Delta] = 10(\lambda_\ominus - \lambda_\oplus) + 12\mu_0 + 30\mu_\ominus - 6\mu_\oplus \tag{1}$$

Proof Table A.1 reports the expected profit from overprovision, the expected profit of honesty, and their difference, when there is certainty and reputation. The first column reports the sequence of realizations of need types for the buyers matched to the expert (1 indicates high need). This sequence will determine when and if a lie is caught, as well as the resulting loss of revenues. Each sequence is equally likely (for simplicity we have dropped the 0.125× term from the next three columns). The second column report expected profits from lying, the third from telling the truth, and the fourth is the their difference.[8] The final row is the sum (expected value ×8) of all the other rows.

Table A.1: Expected Profit of Overprovision versus Honesty by Realization of Type (Certainty)

High	π(**Lie**)	π(**Truth**)	$\boldsymbol{\Delta}\pi$
0,0,0	$\mu_0 B + 2\lambda_\ominus C + 2\mu_\ominus B$	$0 + 2\lambda_\oplus C$	$2(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + 2\mu_\ominus)B$
0,0,1	$\mu_0 B + 2\lambda_\ominus C + 2\mu_\ominus B$	$0 + 2\lambda_\oplus C + \mu_\oplus B$	$2(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + 2\mu_\ominus - \mu_\oplus)B$
0,1,0	$\mu_0 B + 2\lambda_\ominus C + 2\mu_\ominus B$	$0 + 2\lambda_\oplus C + \mu_\oplus B$	$2(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + 2\mu_\ominus - \mu_\oplus)B$
0,1,1	$\mu_0 B + 2\lambda_\ominus C + 2\mu_\ominus B$	$0 + 2\lambda_\oplus C + 2\mu_\oplus B$	$2(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + 2\mu_\ominus - 2\mu_\oplus)B$
1,0,0	$\mu_0 B + (\lambda_\oplus + \lambda_\ominus)C + (\mu_\oplus + \mu_\ominus)B$	$\mu_0 B + 2\lambda_\oplus C$	$(\mu_\ominus - \mu_\oplus)C + (\mu_\oplus + \mu_\ominus)B$
1,0,1	$\mu_0 B + (\lambda_\oplus + \lambda_\ominus)C + (\mu_\oplus + \mu_\ominus)B$	$\mu_0 B + 2\lambda_\oplus C + \mu_\oplus B$	$(\mu_\ominus - \mu_\oplus)C + \mu_\ominus B$
1,1,0	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$\mu_0 B + 2\lambda_\oplus C + \mu_\oplus B$	$\mu_\oplus B$
1,1,1	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	0
$\sum$	$(10\lambda_\ominus + 6\lambda_\oplus)C+$ $(8\mu_0 + 10\mu_\ominus + 6\mu_\oplus)B$	$16\lambda_\oplus C+$ $(4\mu_0 + 8\mu_\oplus)B$	$10(\lambda_\ominus - \lambda_\oplus)C+$ $(4\mu_0 + 10\mu_\ominus - 2\mu_\oplus)B$

Observation 1 In certainty, the profitability of overprovision depends on the experts beliefs about the impact of reputation.

Proof If an expert believes reputation has the maximum impact i.e. experts who have a positive (negative) reputation are always (never) consulted and their advice is always (never) followed, $\lambda_\oplus = \mu_\oplus = 1$ and $\lambda_\ominus = \mu_\ominus = 0$, by Lemma 1 substituting these values into Equation 1 expected net profit of lying is:

[8] For example, in the first row, a lie will be caught in the first round, so the cost of the lie is the difference in the likelihood of being consulted with positive versus negative reputation; while the benefit is one round of the likelihood of the buyer purchasing when there is no reputation rating plus two rounds of the likelihood of the buyer buying when reputation is negative. (In this sequence, the honest expert would never earn profit from the buyer purchasing the service.) Whereas in the sixth row, the lie is not caught until the second round and the dishonest expert incurs the costs for both differences in consulting and buying likelihoods, albeit for only one round.

$$\mathbb{E}[\Delta] = 12\mu_0 - 16$$

Since $0 \leqslant \mu_0 \leqslant 1$, it is unprofitable to lie given any belief about the buying rate in the first round before there have been ratings. However, if the expert believes the consequences of ratings are less than extreme, e.g. $\lambda_\oplus = \mu_\oplus = 0.9$ and $\lambda_\ominus = \mu_\ominus = 0.1$, lying is profitable provided buyers almost always take the first round advice to buy.[9]

$$\mathbb{E}[\Delta] = 10(.8) + 12\mu_0 + 30(.1) - 6(.9) = 12\mu_0 - 10.4$$

Observation 2 If there is no reputation, overprovision is at least as profitable as honesty.

Proof Given that there is no opportunity for reputation, rather than $\lambda_\ominus$ and $\lambda_\oplus$, there is only λ_N. Similarly rather than $\mu_\oplus, \mu_\ominus$ and μ_0, there is only μ_N.[10] Equation 1 simplifies to:

$$\mathbb{E}[\Delta] = 36\mu_N$$

Overprovision is profitable if the probability of buyers purchasing the high-need service when recommended, $\mu_0 > 0$. If $\mu_0 = 0$, the overprovision and honesty are equally profitable.

A.1.2 Diagnostic Uncertainty

In diagnostic uncertainty, the experts receive a noisy signal of buyer type, and services work without fail, so expert dishonesty is NOT the only reason a buyer following expert advice would not get the best outcome possible given their type.

Observation 3 In the Diagnostic Uncertainty Treatment, $\forall\mu_0$, $\mathbb{E}[\Delta] > 0$, i.e. lying is profitable for any belief about μ_0.

Proof Assume a buyer will rate the expert negatively if they buy and do not get the high payout and positively otherwise, but buyers in subsequent rounds then discount the ratings to compensate for the known errors in reputation given the uncertainty. Assuming half the experts lie, only 66% of the experts with a satisfactory rating are actually honest, whereas 33% of those with and an unsatisfactory rating are actually honest.[11] by Lemma 1 substituting $\lambda_\oplus = \mu_\oplus = 0.66$ and $\lambda_\ominus = \mu_\ominus = 0.33$ into Equation 1 expected net profit of lying is:

$$\mathbb{E}[\Delta] = 2.7 + 12\mu_0$$

A.1.3 Service Uncertainty

When the service fails the buyer cannot infer anything about quality of the advice from the expert.

Lemma 2 If there is service uncertainty, the expected net profit from overprovision is,

$$\mathbb{E}[\pi(Lie) - \pi(Truth)] = \mathbb{E}[\Delta] = 64/9\lambda_\ominus - 64/9\lambda_\oplus + 108/9\mu_0 + 24/9\mu_\oplus + 192/9\mu_\ominus \qquad (2)$$

Proof Table A.2 is similar to Table A.1 however is expanded to account for the buyers inability to make inferences when the service fails. There are multiple rows for each sequence. The first two sequences each have three rows. Within each sequence, the first row are expected profits if the expert's dishonesty is caught

[9] Smaller differences between $\mu_\oplus$ and $\mu_\ominus$ make lying profitable with less extreme assumptions about μ_0.

[10] We use μ_N rather than μ_0, because a buyers are more likely to consult an expert who does not have a reputation, but they will be able to rate, than an expert who will never have a rating.

[11] The assumption about the dishonesty rate is higher than we expect, but gives reputation the best shot. When the split is not equal, the base rate drives convergence.

in the first round. The second row is the case of the lie being caught in the second round. The third row is the case that the lie is not caught in the first two rounds. The Pr() column show the probability of each case within the sequence. Each sequence still occurs at equal likelihood, and again we drop those probabilities for simplicity. In later sequences there are fewer rows per sequence, because experts do not lie when high=1, so cannot be detected in these rounds.

Table A.2: Expected Profit of Overprovision versus Honesty by Realization of Type (Service Uncertainty)

High	**Pr()**	π(**Lie**)	π(**Truth**)	$\boldsymbol{\Delta}\pi$
0,0,0	6/9	$\mu_0 B + 2\lambda_\ominus C + 2\mu_\ominus B$	$0 + 2\lambda_\oplus C$	$2(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + 2\mu_\ominus)B$
0,0,0	2/9	$\mu_0 B + (\lambda_\oplus + \lambda_\ominus)C + (\lambda_\oplus + \lambda_\ominus)B$	$0 + 2\lambda_\oplus C$	$(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + \mu_\oplus + \mu_\ominus)B$
0,0,0	1/9	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$0 + 2\lambda_\oplus C$	$(\mu_0 + 2\mu_\oplus)B$
0,0,1	6/9	$\mu_0 B + 2\lambda_\ominus C + 2\mu_\ominus B$	$0 + 2\lambda_\oplus C + \mu_\oplus B$	$2(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + 2\mu_\ominus - \mu_\oplus)B$
0,0,1	2/9	$\mu_0 B + (\lambda_\oplus + \lambda_\ominus)C + (\lambda_\oplus + \lambda_\ominus)B$	$0 + 2\lambda_\oplus C + \mu_\oplus B$	$(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + \mu_\ominus)B$
0,0,1	1/9	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$0 + 2\lambda_\oplus C + \mu_\oplus B$	$(\mu_0 + \mu_\oplus)B$
0,1,0	6/9	$\mu_0 B + 2\lambda_\ominus C + 2\mu_\ominus B$	$0 + 2\lambda_\oplus C + \mu_\oplus B$	$2(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + 2\mu_\ominus - \mu_\oplus)B$
0,1,0	3/9	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$0 + 2\lambda_\oplus C + \mu_\oplus B$	$(\mu_0 + \mu_\oplus)B$
0,1,1	6/9	$\mu_0 B + 2\lambda_\ominus C + 2\mu_\ominus B$	$0 + 2\lambda_\oplus C + 2\mu_\oplus B$	$2(\lambda_\ominus - \lambda_\oplus)C + (\mu_0 + 2\mu_\ominus - 2\mu_\oplus)B$
0,1,1	3/9	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$0 + 2\lambda_\oplus C + 2\mu_\oplus B$	$\mu_0 B$
1,0,0	6/9	$\mu_0 B + (\lambda_\oplus + \lambda_\ominus)C + (\mu_\oplus + \mu_\ominus)B$	$\mu_0 B + 2\lambda_\oplus C$	$(\mu_\ominus - \mu_\oplus)C + (\mu_\oplus + \mu_\ominus)B$
1,0,0	3/9	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$\mu_0 B + 2\lambda_\oplus C$	$2\mu_\oplus B$
1,0,1	6/9	$\mu_0 B + (\lambda_\oplus + \lambda_\ominus)C + (\mu_\oplus + \mu_\ominus)B$	$\mu_0 B + 2\lambda_\oplus C + \mu_\oplus B$	$(\mu_\ominus - \mu_\oplus)C + \mu_\ominus B$
1,0,1	3/9	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$\mu_0 B + 2\lambda_\oplus C + \mu_\oplus B$	$\mu_\oplus B$
1,1,0	1	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$\mu_0 B + 2\lambda_\oplus C + \mu_\oplus B$	$\mu_\oplus B$
1,1,1	1	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	$\mu_0 B + 2\lambda_\oplus C + 2\mu_\oplus B$	0
$\sum$		$(8\mu_0 + 64/9\lambda_\ominus + 80/9\lambda_\oplus)B$ $64/9(\lambda_\ominus + 80/9\lambda_\oplus)C$	$(4\mu_0 + 8\mu_\oplus)B+$ $16\mu_\oplus C$	$(36/9\mu_0 + 8/9\mu_\oplus + 64/9\mu_\ominus)B+$ $64/9(\lambda_\ominus - \lambda_\oplus)C$

Observation 4 Lying is profitable, even if reputation has the maximum impact ($\lambda_\oplus = \mu_\oplus = 1$ and $\lambda_\ominus = \mu_\ominus = 0$), with modest beliefs ($\mu_0 > 0.\overline{370}$) about the buyer's likelihood of taking advice from an unrated expert. If ratings do not work as well, lying will be profitable with even lower values of μ_0.

Proof by Lemma 2 substituting the values above for the λs and μs into Equation 2 expected net profit of lying is:

$$\mathbb{E}[\Delta] = -40/9 + 108/9\mu_0 \geqslant 0 \Rightarrow \mu_0 \geqslant 40/108$$

A.2 Video Instructions

Links to the video instructions for each treatment cell.

Certainty No Reputation	`https://lab.cebex.net/chapman21/videos/VideoA.mp4`
Certainty Reputation	`https://lab.cebex.net/chapman21/videos/VideoB.mp4`
Diagnostic Uncertainty Reputation	`https://lab.cebex.net/chapman21/videos/VideoC.mp4`
Diagnostic Uncertainty No Reputation	`https://lab.cebex.net/chapman21/videos/VideoD.mp4`
Service Uncertainty No Reputation	`https://lab.cebex.net/chapman21/videos/VideoE.mp4`
Service Uncertainty Reputation	`https://lab.cebex.net/chapman21/videos/VideoF.mp4`

A.3 Comprehension Check

All participants needed to pass a comprehension check before participating. They had as many chances as they wanted to answer all the questions correctly. Questions (sets) not answered correctly were highlighted in red. This particular quiz is for an expert in Service Uncertainty. Only Service Uncertainty had *Set 4*. The payouts in the quiz corresponded to those in the subjects treatment. buyers' version of the quiz replaced "the Buyer" with "you". The questions were designed to force participants to look at the payout tables, which were available, and understand that the interests of the two roles are not always aligned.

If a Buyer's Final Payment is 40 ECU which of the following apply? (check one from each set)

Set 1
the Buyer bought Service
the Buyer chose DIY
the Buyer did not consult
the Buyer could have gotten this result either by buying or through DIY
Set 2
the Buyer was a low need buyer
the Buyer was a high need buyer
the Buyer could have been either type
Set 3
the Buyer made the best choice given their type
the Buyer could have made a better purchase choice given their type
the Buyer cannot be certain they made the best choice
Set 4
The option the Buyer chose did not WORK
The option the Buyer chose WORKED

If a Buyer's Final Payment is 1 ECU which of the following apply? (check one from each set)

Set 1
the Buyer bought Service
the Buyer chose DIY
the Buyer did not consult
the Buyer could have gotten this result either by buying or through DIY
Set 2
the Buyer was a low need buyer
the Buyer was a high need buyer
the Buyer could have been either type
Set 3
the Buyer made the best choice given their type
the Buyer could have made a better purchase choice given their type
the Buyer cannot be certain they made the best choice
Set 4
The option the Buyer chose did not WORK
The option the Buyer chose WORKED

If a Buyer's Final Payment is 24 ECU which of the following apply? (check one from each set)

Set 1
the Buyer bought Service
the Buyer chose DIY
the Buyer did not consult
the Buyer could have gotten this result either by buying or through DIY

Set 2
the Buyer was a low need buyer
the Buyer was a high need buyer
the Buyer could have been either type
Set 3
the Buyer made the best choice given their type
the Buyer could have made a better purchase choice given their type
the Buyer cannot be certain they made the best choice
Set 4
The option the Buyer chose did not WORK
The option the Buyer chose WORKED

If a Buyer's Final Payment is 9 ECU which of the following apply? (check one from each set)
Set 1
the Buyer bought Service
the Buyer chose DIY
the Buyer did not consult
the Buyer could have gotten this result either by buying or through DIY
Set 2
the Buyer was a low need buyer
the Buyer was a high need buyer
the Buyer could have been either type
Set 3
the Buyer made the best choice given their type
the Buyer could have made a better purchase choice given their type
the Buyer cannot be certain they made the best choice
Set 4
The option the Buyer chose did not WORK
The option the Buyer chose WORKED

The chance that a buyer in the experiment will have high need and would benefit most from Buying Service is . . .
none (0%)
one quarter (25%)
half (50%)
certain (100%)

A buyer in the experiment will know their own need level before making decision:
True
False

If a buyer intends to choose a "do-it-yourself" (DIY) solution, they are better off not paying to consult the expert:
True
False

A.4 Ratings

Table 5 reports by treatment the number times each rating {☹ Unsatisfied, 😐 Neutral, ☺ Satisfied} was given buy buyers who followed advice.

Table A.3: Ratings Given by Treatment

	Unsatisfied	Neutral	Satisfied
Certainty	5	25	86
Diagnostic Uncertainty	7	25	77
Service Uncertainty	27	18	57

Table A.4 reports χ^2 test for equal distributions of ratings across all treatment combinations.

Table A.4: χ^2 Tests Comparing Rating between Treatments

	Statistic	Diagnostic Uncertainty	Service Uncertainty
Certainty	χ^2	0.61	21.33
	p-value	=0.736	<0.001
Diagnostic Uncertainty	χ^2		15.67
	p-value		<0.001

Table A.5 reports ratings by the ECU's the buyer earned through their decisions for buyers following the experts' advice in the Certainty Treatment.

Table A.5: Ratings Given by ECU in Certainty Treatment Dropping Subjects who Followed Advice

	Unsatisfied	Neutral	Satisfied
11 Low & Buy	4	1	0
19 DIY	0	20	41
27 High & Buy	1	4	45

Table A.6 reports χ^2 test for equal distributions of ratings across all earning combinations within the Certainty Treatment.

Table A.6: χ^2 Tests Comparing Rating between Outcomes

	Statistic	DIY	High & Buy
Low & Buy	χ^2	52.4	35.64
	p-value	<0.001	<0.001
DIY	χ^2		10.87
	p-value		=0.004

Table A.7 reports ratings by the ECU's the buyer earned through their decisions for buyers following the experts' advice in the Diagnostic Uncertainty Treatment.

Table A.7: Ratings Given by ECU in Diagnostic Uncertainty Treatment who Followed Advice

	Unsatisfied	Neutral	Satisfied
11 Low & Buy	5	3	6
19 DIY	1	18	39
27 High & Buy	1	4	32

Table A.8 reports χ^2 test for equal distributions of ratings across all earning combinations within the Diagnostic Uncertainty Treatment.

Table A.8: χ^2 Tests Comparing Rating between Outcomes

	Statistic	DIY	High & Buy
Low & Buy	χ^2	17.07	12.84
	p-value	<0.001	=0.002
DIY	χ^2		5.21
	p-value		=0.074

Table A.9 reports ratings by the ECU's the buyer earned through their decisions for buyers following the experts' advice in the Service Uncertainty Treatment.

Table A.9: Ratings Given by ECU in Service Uncertainty Treatment who Followed Advice

	Unsatisfied	Neutral	Satisfied
1 Buy & Not work	19	3	0
9 DIY & Not work	4	7	6
16 Low & Buy	0	2	1
24 DIY	4	4	20
40 High & Buy	0	2	30

Table A.10 reports χ^2 test for equal distributions of ratings across all earning combinations within the Service Uncertainty Treatment.

Table A.10: χ^2 Tests Comparing Rating between Outcomes

	Statistic	DIY & Not Work	Low & Buy	DIY	High & Buy
Buy & Not Work	χ^2	17.02	13.64	29.63	49.03
	p-value	<0.001	=0.001	<0.001	<0.001
DIY & Not Work	χ^2		1.08	6.03	20.07
	p-value		=0.584	=0.049	<0.001
Low & Buy	χ^2			4.85	4.82
	p-value			=0.088	=0.028
DIY	χ^2				6.43
	p-value				=0.04

Table A.11 reports the coefficient estimates and cut points from ordered probit regressions on the ratings given by buyers who follow experts' advice. Lower values indicate less favorable ratings. Errors are clustered on buyers. The first model tests how treatment interacts with expert over-provision to impact ratings. Over-provision is penalized but there are smaller penalties (coefficients estimates) when over-provision is interacted expert diagnostic and none in service uncertainty treatments, than the Certainty Treatment. There are also lower ratings in service uncertainty, even when experts are truthful. The second specification adds a variable to indicate that the selected option failed to work. This variable is negative, showing that penalize the expert when the selection does not work, despite that this was random and independent of the expert. This explains the negative rating for truthfulness in service uncertainty. This confirms graphically what Figure 8 shows.

Table A.11: Coefficients Estimates from Ordered Probit Regression on Rating

	(1) Rating	(2) Rating
Treatment x Advice:		
Cert Rep × Over-Provision	-2.666***	-2.861***
	(0.706)	(0.791)
DiagUncert Rep × Truthful	-0.0926	-0.0961
	(0.198)	(0.211)
DiagUncert Rep × Over-Provision	-1.209**	-1.312**
	(0.445)	(0.488)
ServUncert Rep × Truthful	-0.871***	0.0828
	(0.195)	(0.270)
ServUncert Rep × Over-Provision	-0.901*	-0.954*
	(0.414)	(0.408)
ProdFail		-2.238***
		(0.381)
cut1		
Constant	-1.757***	-1.988***
	(0.171)	(0.212)
cut2		
Constant	-0.865***	-0.854***
	(0.148)	(0.162)
sigma2_u		
Constant	0.115	0.185
	(0.138)	(0.195)
Observations	327	327
Number of Buyers	152	152
Log pseudolikelihood	-252.3	-220.5

Clustered robust Std. Err. in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.12 reports marginal effect from a model without the *Service Fails* variable.

Table A.12: Estimated Marginal Effect from Ordered Probit Regression on Rating

	☹	😐	☺
Cert Rep Over-Provision	0.757***	-0.00770	-0.750***
	(0.168)	(0.104)	(0.0743)
DiagUncert Truthful	0.00942	0.0165	-0.0259
	(0.0203)	(0.0350)	(0.0553)
DiagUncert Over-Provision	0.254	0.168***	-0.422**
	(0.139)	(0.0283)	(0.153)
ServUncert Truthful	0.153***	0.143***	-0.296***
	(0.0409)	(0.0287)	(0.0626)
ServUncert Over-Provision	0.161	0.147**	-0.307*
	(0.105)	(0.0495)	(0.151)
Observations	327	327	327

Clustered robust Std. Err. in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$