8-26-2021

Uncertainty and Reputation Effects in Credence Goods Markets

Eric Schniter
Chapman University, schniter@chapman.edu

J. Dustin Tracy
Chapman University, tracy@chapman.edu

Vojtěch Zíka
Chapman University

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/

Uncertainty and Reputation Effects in Credence Goods Markets

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

April 26, 2024

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. 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 buyers do not always work. Both sources of uncertainty make detection of expert dishonesty more difficult, so are hypothesized to increase dishonesty by experts and decrease buyers' 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 dishonesty and attenuate distrust by creating reliable reputations. In contrast to hypotheses, we find that uncertainty has no effect on honesty and increases trust; additionally, 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 overprovided (not in the buyer's best interest).

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

JEL-Codes: *D82, L14, L15*

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 eschewed. However, in other

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.

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

[†]Department of Health Management, Economics and Policy, School of Public Health, Augusta University & Economics Science Institute, Chapman University; jtracy@augusta.edu

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

[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.

situations, where consumers follow expert advice and get the recommended treatment, they find it difficult to judge whether the expert was honest about which treatment was best. 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* (Darby and Karni, 1973), include medical care, car and computer repairs, financial advice, and taxi rides in unfamiliar places. There have been many laboratory and field experiments regarding credence goods, which we review in Section 2.

Notably, these experiments have focused on situations in which uncertainty is not present (Kerschbamer and Sutter, 2017). Throughout the paper, we reserve the term uncertainty, for the standard economic definition, i.e., for any probabilistic process for which the outcome has yet to be determined or revealed, so no one knows the realized value, though some or all parties know the probabilities of the various outcomes (Neumann, 1944). This is distinct from situations in which a consumer might be unsure about which option is best, hence consult an expert with greater knowledge. Or as Arrow (1963, p. 946) explains, "information, in the form of skilled care, is precisely what is being bought from most physicians, and, indeed, from most professionals." By starting with experiments investigating decisions under certainty, researchers gained insights into the dynamics of these markets. The literature shows that reputation paired with competition and liability (recourse when the service is not adequate) increases honesty and trust, when there is certainty. However, credence-good markets outside of experiments rarely involve simple, certain solutions. "That risk and uncertainty are, in fact, significant elements in medical care hardly needs argument" (Arrow, 1963, p. 946). We contend this extends to many other credence goods. So, 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, so is misleading, albeit unintentionally and through no fault of their own; e.g., a doctor makes a diagnosis based on a 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. Our service uncertainty is identical to what (Arrow, 1963, p. 951) described as product uncertainty, explaining "[u]ncertainty as to the quality of the product is perhaps more intense here than in any other important commodity. Recovery from disease is as unpredictable as is its incidence." Both types of uncertainty create 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) report on experiments contemporaneous to our with diagnostic uncertainty, including a treatment in which the accuracy of the diagnosis is endogenous. To our knowledge, we are the first experiment with stochastic outcomes for credence goods, despite Arrow's identification of product uncertainty a half-century ago.

We make an additional contribution to the literature by implementing a ratings-based reputation mechanism, based on those popular in trust-based exchange markets outside the laboratory. Most 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. One exception is Mimra et al. (2016) who have a reputation treatment in which there is a public history of incidents of undertreatment. Similar to a contemporaneous experiment conducted by Angerer et al. (2021), our design and treatments involve having buyers rate the experts they consult, with these ratings being accessible to future buyers. Whereas Angerer et al.'s buyers use ratings to make a choice among four experts, our buyers can use ratings to decide whether or not to consult the expert they have been randomly matched with that round. We allow three ratings: "Satisfied", "Neutral" or "Unsatisfied", which is much like the rating system on eBay.

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.[2] 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 that among subjects in a trust game, those in the "reputation-stranger" treatment were more trusting and more trustworthy than subjects in the stranger treatment and less trusting and trustworthy than those 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 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 between-subject design, in which we varied whether experts had reputations and each source of uncertainty, but did not run 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 (rather than always). In the Reputation Treatment, 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". The rating was made available to the buyers who encountered that expert in later rounds, prior to the buyers making a decision. 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, that without uncertainty, experts were more likely to provide dishonest advice when there were no ratings. However, surprisingly, when there was uncertainty, rates of dishonest advice were no higher in treatments where consumers did not rate experts. Equally surprising is that the lack of ratings did not create any additional 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 was more frequent, despite that their advice was less informative. In the diagnostic uncertainty conditions, buyers seemed to give experts the benefit of the doubt; satisfied was the most common rating when low-need buyers selected high-need service. In the Service Uncertainty Treatment, 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 decrease consultation, which led to efficiency loss. Our results suggest, that when there is uncertainty, ratings do not improve outcomes and may worsen them.

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

2 Background

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) provide the foundation of modern theory regarding asymmetric information. Arrow, in particular, calls attention to doctors' conflicting

[2]The findings of Angerer et al. (2021) also suggest this. While they do not have a fixed pair treatment, they find that relative to no reputation both private and public ratings significantly decrease the dishonesty of experts who compete for buyers. The public rating system performs slightly better than the private system.

interests between prescribing treatment that is best for the patients' health and that is most profitable. While he notes ethical considerations might restrict the physician from exploiting asymmetrical information, he also argues "The elusive character of information as a commodity suggests that it departs considerably from the usual marketability assumptions about commodities" (p. 946).

Much work has been done to formalize and expand the theory on asymmetric information, showing how experts can exploit this asymmetry. 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, though experience goods may entail information asymmetry before they are experienced, after normal use there is no information asymmetry; e.g., 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 the 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 credence-goods experiments use this binary model. Pitchik and Schotter also argue an increase in experts' diagnostic accuracy (expert certainty in our terminology) will lead to a decrease in honesty, though this depends on heterogeneity of expert certainty. Wolinsky (1993) posits a model in which expert diagnosis is imperfect. 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 with acquaintances, such as offered by review and rating systems. Bester and Dahm (2018) assume diagnosis requires costly and unobservable effort, which contributes 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.

Our review of the related experimental literature on dishonest advice starts with papers not always cited in when reviewing credence-goods literature, despite the evidence they provide about when asymmetric information is likely to be exploited. Plott and Wilde (1982) present evidence that when buyers have uncertainty regarding their needs, 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 c. 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 on 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 the profitability of lying, that the authors attributed to subject error.

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). Laboratory experiments have tested mechanisms for restraining expert dishonesty. Dulleck et al. (2011) report on a credence-goods experiment with 16 treatment cells that vary whether there is reputation (if sellers have identities rather than being anonymous), competition (if the buyer can choose from multiple sellers), liability (if the buyer has recourse when underprovision), and verifiability (if overcharging is possible). They find evidence of overcharging, overprovision, and underprovision, concluding that verifiability does little to improve efficiency but liability increases efficiency. They also report that reputation alone has little impact aside from when liability and verifiability are absent. Kerschbamer et al. (2017) present evidence suggesting that verifiability does not increase efficiency because of the underlying heterogeneity in social preferences. 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 in the role of experts. In an experiment testing how payment systems influence physicians' trade-off between costly effort and patient benefit, Martinsson and Persson (2019) find that neither risk nor ambiguity affect choices.

Field experiments confirm many but all of the results from laboratory experiments and expand our understanding of the problems and viability of solutions for efficient credence-goods markets. In a field experiment, concerning auto mechanics, Schneider (2012) finds significant levels of dishonesty in the form of overprovision and underprovision. While the finding of underprovision is in contrast with theory and not found in laboratory experiments, We note the underprovision, resulted in only small losses of revenue for the mechanics. Surprisingly, Schneider finds that the suggestion of repeated business did not significantly improve recommendations. The study sampled relatively few repair shops, it is not clear whether the mechanics believed overprovision or underprovision would likely be detected. 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 are more likely to overcharge the customer. In a related field experiment, Balafoutas et al. (2017) find that "second-degree moral hazard"—situations in which the buyer will be reimbursed for the charges—also increase overcharging. Likewise, Kerschbamer et al. (2016) find that insurance coverage, which reduces the uncertainty of potential loss for the customer, induces dishonesty of sellers in credence-goods markets. Gottschalk et al. (2020) find that 28% of Swiss dentists recommended unnecessary fillings; lower income of patients, shorter waiting times for appointments, and ownership of the practice were associated with increased likelihood of overprovision.

In credence-goods experiments, reputation is generally induced through repeated interactions and personal history. However, outside these experiments, reputation is comprised of what consumers formally and informally share with each other about interactions with a seller or expert, either in interpersonal exchanges or through rating systems and similar mechanisms. There is considerable research showing that consumer ratings of sellers can identify low-quality products and discipline dishonest sellers. Livingston (2005) finds that positive reviews of eBay sellers increase the sale price and probability that items they auction receive bids. Houser and Wooders (2006) confirm the price effect, but do not find any effect for buyer reputation. Lucking-Reiley et al. (2007) find the effect of negative reviews is much stronger than positive reviews. Cabral and Hortaçsu (2010) find that negative feedback on eBay can have a deleterious impact on a seller's future sales. Huck and Lünser (2010) allow consumers to share information with each other about past experiences with sellers while varying network structure and group size. Their findings indicate partial information can be very effective at instilling trust and improving efficiency. Luca (2016) finds that a one-star increase in a seller's average Yelp review is associated with a 9% increase in revenue. Kerschbamer et al. (2019) find that better-rated computer shops (according to Google and Yelp! reviews) charged less for the same repair. They describe the finding as "striking" because all the shops were able to repair the computer, implying that they view the rating as an indication of quality. We suggest that if ratings are viewed as an indication of value (quality given the price), the result is much more intuitive. Huesmann et al. (2020) show that ratings systems that lack granularity can reduce the quality of service. Reimers and Waldfogel (2021) show that consumer reviews, especially positive reviews, increase total sales and improve consumer welfare, suggesting that the availability of reviews increases market participation. In a health-care framed laboratory experiment, Angerer et al. (2021) find that either public or private ratings combined with competition reduce overprovision and overcharging in credence-goods markets.

However, there is some reason to question how well consumer ratings will work in credence-good markets,

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

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. The authors suggest two mechanisms, through which disclosure could improve ratings. The first is that when patients have a bad outcome, they give doctors the benefit of the doubt — despite doctor disclosure. Another possibility is that when doctors disclose uncertainties and patients go on to have positive outcomes, they experience relief that things did not result badly and this increases their satisfaction. In contrast, without disclosure patients do not experience the surprise of discovering and anxiety of considering the uncertainties, nor could they experience the extra satisfaction of relief from those costly emotions. 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 bad news and is particularly strong when the news comes as a surprise and requires the recipient to work out internally inconsistent beliefs or experience negative consequences. Filippas et al. (2022) 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 experimental 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 implement an additional source of uncertainty common in credence-goods, service uncertainty, per Arrow (1963). Even if a doctor accurately diagnoses a patient and prescribes appropriate treatment, the treatment may not work for that particular patient. Similarly, a financial adviser could make sound 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. We focus on overprovision. There is no way our experts can overcharge. While, we allow for underprovision, it decreases expert earnings so we do not expect it. Our environment, like those of Plott and Wilde (1982) and Dulleck et al. (2011), includes a market entry decision (i.e., whether to consult the expert for advice), and allows buyers to disregard expert advice. In our design, the low-need good is a "do it yourself" (DIY) solution which is available to the consumer independent of the expert. For example, in the case of a sprained ankle, the expert medical advice might be to ice it, wrap it in an elastic bandage, and not put weight on it for the recovery period. Consumers, who are skeptical of medical experts, might pursue this treatment without consulting an expert. However, if the injury is more serious (e.g. a torn ligament) and actually requires surgery, the consumer would need to consult with a medical expert before being able to elect surgery. In such cases, skeptical consumers who forgo consultation are worse off because of their skepticism, so more trusting consumers gain peace of mind through consulting, even if the eventual treatment is DIY. While "commitment" to following advice has been identified as important to market efficiency (Dulleck and Kerschbamer, 2006), we opted for an environment in which we had both consultation and advice-following measures of consumer trust.

2.1 Model

Our setup is typical of credence-good experiments. Subjects are assigned to one of two roles, either "buyer" or "expert", which they maintain throughout the experiment. Each round, buyers are assigned a need type {high, low}; with probability $0 < p < 1$ they are high need.[4] Their type is not revealed to them, but buyers can seek an expert's advice. If consulted, experts receive a signal with accuracy ($0.5 < s \leqslant 1$) about

[4] p is bound so the game is not degenerate with the same need type for all buyers.

the buyer's type.[5] While in previous experiments, the s parameter has typically remained fixed at 1, we exogenously manipulate its value in our study.[6] Experts then advise buyers, on which service would best serve them. There are two types of services, both work with probability q. The less expensive service, if it works fulfills only the low need, while the more expensive service fulfills both needs, if it works. If it does not work, neither service fulfills any need. Net of cost, low-need buyers benefit more from the former while high-need buyers benefit more from the latter. Another innovation of our model is $q < 1$.[7] 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; however, if the low-need buyer purchases the high-cost service, they pay the cost but do not receive the benefit the high-need consumer does, so the expert's dishonesty is discovered. Therefore, the experts balance the additional profit from overproviding against the potential cost of lost future profits if their dishonesty is discovered.

In short, we expand the traditional setup by adding two sources of uncertainty. The first is diagnostic uncertainty; experts receive a noisy signal of buyer type ($s < 1$). The other is service uncertainty; the DIY and expert services do not always work ($q < 1$).

Regardless of either uncertainty, expected expert pay for honest advice is:

$$\mathbb{E}_t[\pi] = pB\mu(R_t) + \mathbb{E}_t\left[\sum_{\tau=t+1}^{T} \lambda(R_\tau)\Big(C + pB\mu(R_\tau)\Big)\right]$$

where B is the expert's profit from providing the high-need service (relative to proving the low-need service), C is the expert's profit from being consulted (or providing the low-need service), λ is the probability that the buyer consults the expert, and μ is the probability that the buyer follows the expert's advice. Both depend upon the expert's reputation, R_τ at time τ.

Expected pay for overprovision is:

$$\mathbb{E}_t[\pi] = B\mu(R_t) + \mathbb{E}_t\left[\sum_{\tau=t+1}^{T} \lambda(R_\tau)\Big(C + B\mu(R_\tau)\Big)\right]$$

The difference is:

$$\Delta\,\mathbb{E}_t[\pi] = (1-p)B\mu(R_t) + \mathbb{E}_t\left[\sum_{\tau=t+1}^{T} \lambda(\Delta R_\tau)\Big(C + (1-p)B\mu(\Delta R_\tau)\Big)\right] \tag{1}$$

where ΔR_τ is the expected change in reputation due to overprovision. Reputation is normalized to the range (0,1), where 1 is flawless. With each additional interaction expert i's reputation evolves:

$$\begin{aligned}\mathbb{E}[R_{i,t+1}] = \alpha R_{i,t} + (1-\alpha)\Big[&\theta\Big(p + (1-p)(1-q) + (1-OP_i)(1-p)sq\Big) \\ &+\phi\Big((1-p)(1-s)q + OP_i(1-p)sq\Big)\Big]\end{aligned} \tag{2}$$

where OP_i is the propensity of the expert to overprovide, i.e. intentionally advise low-need buyers to purchase the high-need good; α, (0,1), is parameter capturing persistence; and θ and ϕ are the respective intensities (or probabilities) of reward and punishment. If buyers are unwilling to rate experts as highly or harshly when there is a noisy signal (compared to when the signal is pure), $1 > \theta > \phi > 0$. If there is no adjustment in the ratings for the noise then, $\theta = 1$ and $\phi = 0$. The bracket contains new information. Within it, the first term is good ratings, and the second is bad ratings. We assume better reputation cannot decrease the probability of consultation and advice following: $\lambda(R)' \geqslant 0$ and $\mu(R)' \geqslant 0$.

If $q = 1$ and $s = 1$, Equation 2 simplifies to:

[5] A signal with an accuracy of 0.5 contains no information.

[6] Liu et al. (2019) allow the expert to exert costly effort which determines the accuracy. Balafoutas et al. (2020) have a treatments in which s is endogenous and in which $s = 0.7$.

[7] We do not vary q across the goods. If we were to vary the probability, we would have $q_H \neq q_L$.

$$\mathbb{E}[R_{i,t+1}] = \alpha R_{i,t} + (1-\alpha)\left[\theta\Big(p + (1-OP_i)(1-p)\Big) + \phi OP_i(1-p)\right]$$

Here we expect $\theta = 1$ and $\phi = 0$, as OP_i is a clear indicator of expert honesty. We include them for consistency with later equations. If there is no overprovision, $OP_i = 0$, and no adjustment, $\theta = 1$, the bracketed term is 1, entirely good reputation. However, if there is overprovision $OP_i = 1$ and no adjustment $\phi = 0$, the term is p, the expert receives positive ratings only from high-need consumers.

If $q = 1$ and $s < 1$, Equation 2 simplifies to Equation 3:

$$\begin{aligned}\mathbb{E}[R_{i,t+1}] = \alpha R_{i,t} + (1-\alpha)\Big[\theta p(1 - \Big(sOP_i + (1-s)(1-OP_i)\Big) + \\ \Big(sOP_i + (1-s)(1-OP_i)\Big)\Big(1 - \phi(1-p)\Big)\Big]\end{aligned} \tag{3}$$

If $q < 1$ and $s = 1$, Equation 2 simplifies to Equation 4:

$$\begin{aligned}\mathbb{E}[R_{i,t+1}] = \alpha R_{i,t} + (1-\alpha)\Big[\theta\Big(p + (1-p)(1-q) + (1-OP_i)(1-p)q\Big) + \\ \phi OP_i(1-p)q]\end{aligned} \tag{4}$$

Note $(1-p)(1-q)$ of the ratings become favorable (have moved from the second to the first term in the bracket) because failure of the service obscures overprovision.

Proposition 1: Without reputation payoffs do not provide incentives for honesty.

Proof: Without reputation ΔR_τ is undefined, so Equation 1 reduces to the first term, which cannot be negative.

In other words, overprovision is expected to increase profit, when there is no reputation, as such we expect more overprovision in this treatment. We also expect buyers to anticipate this overprovision and to be more wary of experts.[8]

Hypothesis 1: The Reputation Treatment (relative to no reputation) will increase the experts' honesty and the buyers' trust.

To test this hypothesis we will analyze the data for evidence of: a) decreased rates of overprovision, b) increased rates of consultation, c) increased rates of buyers selecting the high-need service when advised to. Additionally, we expect: d) increased buyers' welfare (earnings).[9]

Proposition 2: An increase in the signal quality does not increase the expected reputation (does not decrease cost) and the payoff of overprovision.

Proof: (from Eq. 3)

$$\frac{\partial\, \mathbb{E}[R_{i,t}]}{\partial OP_i} = (1-\alpha)\left[-\theta(1-p)s + \phi(1-p)s\right]$$

$$\frac{\partial^2\, \mathbb{E}[R_{i,t}]}{\partial s \partial OP_i} = (1-\alpha)\left[-\theta(1-p) + \phi(1-p)\right]$$

[8]Because our experts are in competition against the DIY option, we expect results closer to Dulleck et al.'s (2011) Competition/Reputation Treatment than to their Reputation Treatment. Schneider (2012) also fails to find an effect from reputation. However, the study had relatively few experts and it was not clear what they believed about the probability of detection of overprovision.

[9]All Hypotheses are preregistered at `https://aspredicted.org/blind.php?x=tp4az6`.

$0 \leqslant \phi < \theta \leqslant 1,\ p < 1,\ \alpha < 1$

$$\frac{\partial^2\, \mathbb{E}[R_{i,t}]}{\partial s \partial OP_i} \leqslant 0$$

if we restrict $p \geqslant 0.5$[10] $\lambda(R)' \geqslant 0$ and $\mu(R)' \geqslant 0$

$$\frac{\partial \Delta\, \mathbb{E}[\pi]}{\partial s} \leqslant 0$$

Hypothesis 2: Within the Reputation Treatments, the Diagnostic Uncertainty Treatment ($s < 1$) will (relative to certainty ($s = 1$)) decrease the experts' honesty and the buyers' trust.[11]

To test this hypothesis we will analyze the data for evidence of: a) increased rates of overprovision, b) decreased rates of consultation, c) decreased rates of buyers selecting the high-need service when advised to, and d) decreased likelihood of consultation or advice following conditional on rating seen. Additionally, we expect: e) a reduction in the frequency of experts who overprovide receiving "unsatisified" ratings, and f) decreased buyers' welfare (earnings).

Proposition 3: An increase in the probability that service works does not increase expected reputation (does not decrease reputation cost) and the payoff of overprovision.

Proof: (from Eq. 4)

$$\frac{\partial\, \mathbb{E}[R_{i,t}]}{\partial OP_i} = (1-\alpha)\left[-\theta(1-p)q + \phi(1-p)q\right]$$

$$\frac{\partial^2\, \mathbb{E}[R_{i,t}]}{\partial q \partial OP_i} = (1-\alpha)\left[-\theta(1-p) + \phi(1-p)\right]$$

$$\frac{\partial^2\, \mathbb{E}[R_{i,t}]}{\partial q \partial OP_i} \leqslant 0$$

$$\frac{\partial \Delta\, \mathbb{E}[\pi]}{\partial q} \leqslant 0$$

Hypothesis 3: Within the Reputation Treatments, the Service Uncertainty Treatment ($q < 1$), relative to certainty ($q = 1$), will decrease the experts' honesty and the buyers' trust.[12]

To test this hypothesis we will analyze the data for evidence of: a) increased rates of overprovision, b) decreased rates of consultation, c) decreased rates of buyers selecting the high-need service when advised to, and d) decreased likelihood of consultation or advice following conditional on rating seen. Additionally, we expect it to: e) reduced frequency of experts who overprovide receiving "unsatisified" ratings, and f) decreased buyers' welfare (earnings).

Appendix A.1 includes two additional propositions, which we do not test in this paper.

[10] If $p < 0.5$, then honest experts unintentionally recommend increased purchase of the high-need good and thereby decreasing the relative profit of intentional overprovision, at least for the first (immediate) term; the impact on the second term is negative, so the overall impact is ambiguous.

[15] See previous footnote.

3 Design

Our experiment is between subjects and has a 2x2x2 treatment design (though we only run 6 of 8 possible treatments). There are two uncertainty treatments: Diagnostic Uncertainty and Service Uncertainty; and an (expert) Reputation Treatment. We do not interact the two uncertainty treatments because theory does not suggest that the effects would increase or decrease in interaction. In the Diagnostic Uncertainty Treatment, subjects were told the expert administered a test that was 75% accurate, whereas it was 100% accurate in the Certainty (baseline) Treatment. In the Service Uncertainty Treatment, the DIY and expert services only work $66.\overline{6}\%$ of the time. (They work 100% of the time in baseline.) Appendix A.2 quantifies our hypotheses for our particular parameters. We show that if a negative reputation has the maximum cost, i.e., buyers always consulted and followed the advice of positively rated experts and never consulted negatively rated experts, overprovision would not be profitable in expectation when there was certainty. When there is uncertainty, it follows that reputation will have a lower cost, thus overprovision is profitable.

In the Reputation Treatment, buyers rate their experience with an expert, and that rating {☺ Satisfied, 😐 Neutral, ☹ Unsatisfied} is seen by the next buyer matched with that expert. Thus a buyer who has been the victim of overprovision (having followed 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. We used a three-tier system, because it allows for a neutral rating and simplifies the interpretation and analysis. All feasible ratings are equal in a non-optional effort cost, i.e., we make the buyers pick a rating, so do not see why they would not click on the one that honestly reflected their experience. As the game's rewards do not offer any monetary gains or losses for buyers to provide accurate or inaccurate ratings, we believe the setup creates inherent motivations for accurate ratings. Our experimental design allows us to investigate if these intrinsic incentives impact ratings: the desire to punish experts who overprovide will lead buyers to negatively rate those experts; and an aversion to punishing expert who provide honest advice will result in positive ratings. Additionally, we note that if buyers did not accurately rate experts, then they should expect other buyers to also inaccurately rate experts, so should disregard ratings when making decisions. The warm glow from contributing to a public good (Andreoni, 1990) or other regarding preferences (Fehr and Schmidt, 1999) might also motivate buyers to provide accurate ratings to their fellow buyers.

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 subjects who could read a brief description of the study and then opt to proceed to the study's web-based software. Table A.4 reports subject characteristics. Our software was programmed 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 subjects received instructions (based on feedback in a classroom pilot we included an option to watch a short, narrated, video version (see Appendix A.6.3) of the instructions). Subjects took a quiz (see appendix A.7) to test comprehension before participating in the experiment. They had to complete the quiz without error to advance to the study. If a subject 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). Subjects, who found the quiz too tedious, could choose not to participate and perhaps return to Prolific to find another paid study.

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 straightforward, aided in online implementation, and simplified 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 buyers with their diagnostic test results. In the Uncertainty Treatment, there was a random draw to determine if their diagnostic test accurately determined their need type. After seeing the expert's advice, buyers made the choice as to whether to buy the expert's service (or to use the DIY service). Buyers learned the result of their decision, and then if they consulted the expert that round and were in the Reputation Treatment, they rated the expert {☹ Unsatisfied, 😐 Neutral, ☺ Satisfied}. 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 post-experiment class discussion. Students indicated a desire to see the information in both formats.

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

Figure 2 is a screenshot of the payout tables for the Service Uncertainty Treatment. 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 overprovision.

Figure 2: Payouts tables for the Service Uncertainty Treatment

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. 27^{th} to Feb. 4^{th}, 2021 on Prolific.co. There were 297 experts and 324 buyers.[13] Experts earned \$5.70 and buyers \$3.78,[14] 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.

We begin with a summary of our findings with regard to our hypotheses, in order to provide the reader with context for the analyses which substantiate these findings, prior to discussing the details of those analyses.

Finding 1: Consistent with our hypothesis, in the Reputation Treatment, experts were more honest. However, contrary to our hypothesis, buyers were not more trusting.

[13] The software was programmed to stop admitting buyers once the requisite number of buyers completed the experiment, some extra buyers were admitted between the time the requisite buyers started and completed the experiment.

[14] We intended earnings to be roughly equal; however, consultation and buying rates were higher than expected.

Finding 2: Contrary to our hypothesis, in the Diagnostic Uncertainty Treatment, experts were no less honest, nor were buyers less trusting.

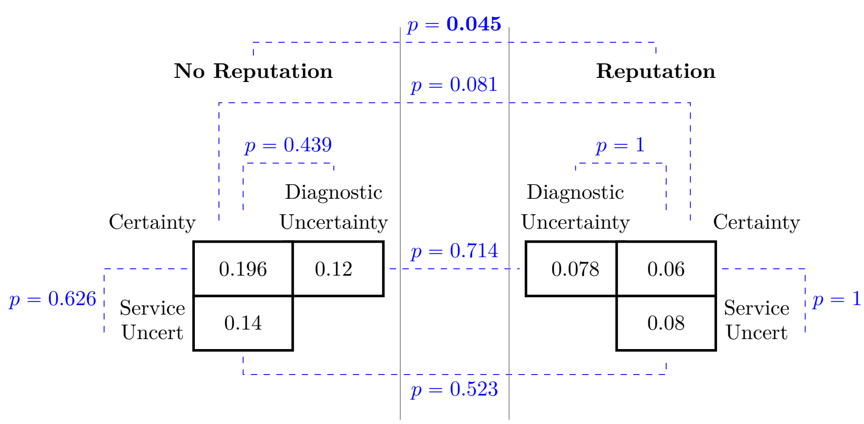
Finding 3: Contrary to our hypothesis, in the Service Uncertainty Treatment, experts were no less honest, nor were buyers less trusting.

Recall to test hypotheses, we sought evidence from a) rates of overprovision, b) rates of consultation, c) rates buyers selected the high-need service when advised to, and d) rates of consultation or advice following conditional on rating seen. The remainder of this section analyzes each of the above in turn. In addition, we also analyze determinants of ratings (before d) so that it proceeds our analysis conditioning on ratings. Finally, we examine earnings of both buyers and experts in each treatment.

4.1 Expert Decisions

Experts made two decisions: the advice to give to buyers whose test indicates low need, and the advice to give to buyers whose test indicates high need. 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 subject 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. Despite that it theory and practice overprovision results in higher pay (see Appendix A.2), we see relatively low rates of overprovision in all treatments. 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 corresponding tests are beside lines.[15]

Figure 3: Rates of Overprovision by Treatment



In support of Finding 1, as hypothesized, the Reputation Treatment, in which buyers rate experts and can avoid experts with negative ratings, has lower rates of overprovision than the No Reputation Treatment. 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 $p < 0.05$. In support of Finding 2, 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. In support of Finding 3, contrary to our hypothesis, differences in rates of overprovision are not statistically significant between service

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

uncertainty (where when the service does not work overprovision is undetectable) and certainty. Rates are actually lower when there is no reputation.

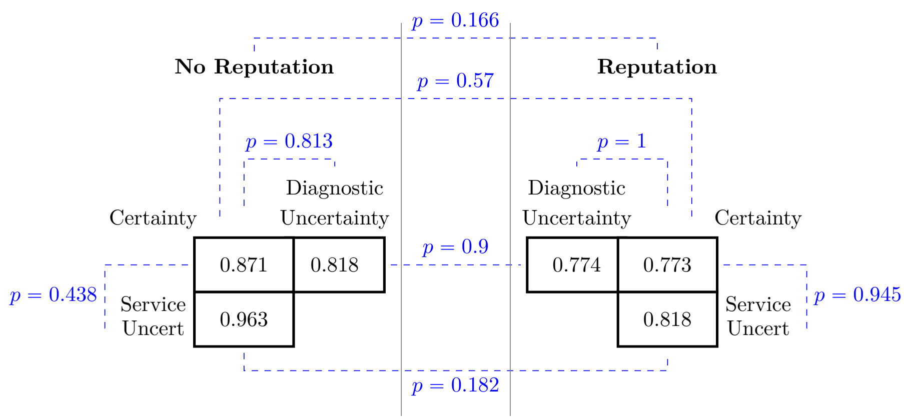
4.2 Buyer Decisions

Buyers made up to three decisions each round. In the first round, buyers must consult and make 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 were no ratings. Dashed lines indicate χ^2-test; p-values for the corresponding tests are beside lines.

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



Within the first round, contrary to our hypothesis, the Reputation Treatment, in which experts had greater incentives to be honest, has lower rates of accepting advice to buy than the No Reputation Treatment. The p-value of a χ^2-test pooling all three cells is 0.166 and not statistically significant. Similarly, the difference across individual cells is not statistically significant. These results support the second part of Finding 1. Within the first round, contrary to our hypothesis, differences in rates of accepting advice to buy are not statistically significant between certainty and diagnostic uncertainty (where the incentives to be honest were weakened by the fact deliberate overprovision cannot be identified). These results support the second part of Finding 2. Within the first round, contrary to our hypothesis, differences in rates of accepting advice to buy are not statistically significant between certainty and service uncertainty (where the incentives to be honest were weakened by the fact that when the service does not work it obscures overprovision). These results support the second part of Finding 3.

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 ratings from buyers who did not follow advice because, the relationship between outcome and rating given should be substantially different depending upon whether the advice was followed, perhaps being inverse. Brackets with p-values between bars indicate χ^2 tests of ratings across treatments. 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.6.

Table A.5 reports counts. To understand what is driving the treatment differences, we examine how ratings varied by outcome within each treatment.

Figure 5: Rating by Treatment

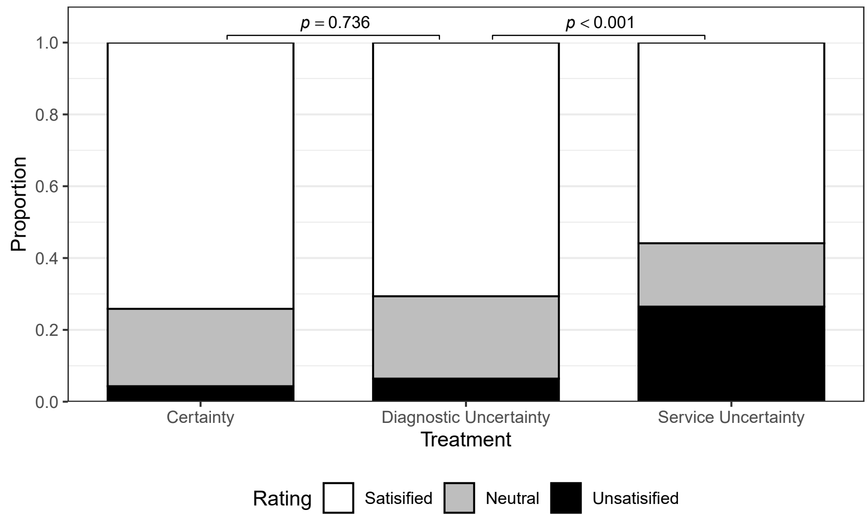


Figure 6 displays ratings given by buyers following the experts' advice in the Certainty Treatment. Each bar represents a particular level of earnings (Bottom Axis), and shows proportions of each of the three ratings within that level. Advice and type (if the outcome reveals it) are on the Top Axis. The ratings, given the outcomes, are much as expected. The majority of buyers, who bought when it was not in their best interest, and earned only 11 ECU, rate the experts, who advised them to do so, negatively. About a fifth of buyers in this situation rate the experience as neutral; this may indicate confusion or a hesitancy to punish the experts, who overprovide. The majority of buyers with both of the other outcomes rate the experts who gave them honest advice positively. It is notable that almost a third of the buyers who were advised to and chose DIY gave neutral ratings, while less than a tenth do not gave satisfied ratings for High & Buy. The fact that DIY has more neutral ratings than High & Buy evidences that buyer were more likely to base ratings on a 'one-step' logic rather than a more informative 'two-step' logic. Advising DIY is strong evidence that the expert does not overprovide, but requires the buyer to realize two things. 1) If an expert were overproviding they would never advise DIY, because they advise the high-need buyer to buy (underproviding is costly to the expert); and also, by definition, advise the low-need buyers to buy the high-need service. 2) Because this expert advised DIY and overproviding experts never advise DIY, this expert must not overprovide. On the other hand, High & Buy provides no information; both honest and dishonest experts would offer the same advice. All the buyer can glean is this instance of advising buy was not overprovision, but cannot infer anything about other instances of advising to buy. Brackets with p-values between bars indicate χ^2 tests of ratings across outcomes. Chi values and the test of the final combination are reported in Table A.8. Table A.7 reports counts.

Figure 7 displays ratings given by buyers following the experts' advice in the Diagnostic Uncertainty Treatment. Each bar represents a particular level of earnings (Bottom Axis) and shows the proportions of each of the three ratings within that level. Advice and type (if the outcome reveals it) are on the Top Axis. Buyers who received only 11 ECUs appear to gave experts the benefit of the 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. Brackets with p-values between bars indicate χ^2 tests of ratings across outcomes. Chi values and the test of the final combination are reported in Table A.10. Table A.9 reports counts.

Figure 6: Rating by Outcome in Certainty Treatment

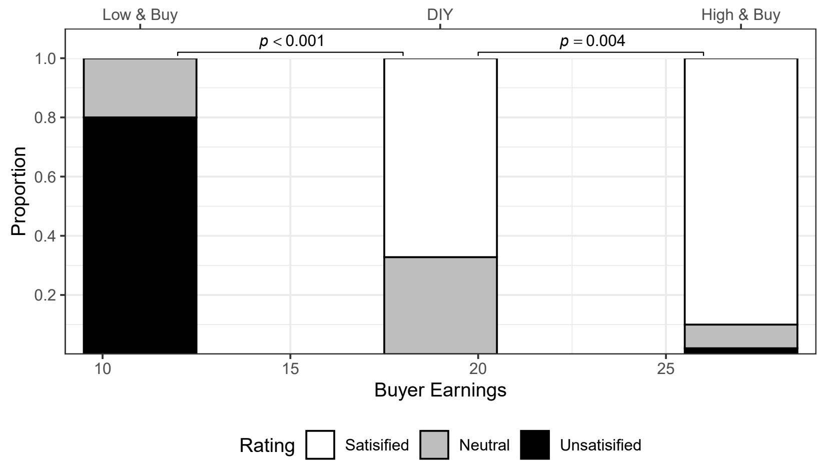


Figure 7: Rating by Outcome in Diagnostic Uncertainty Treatment

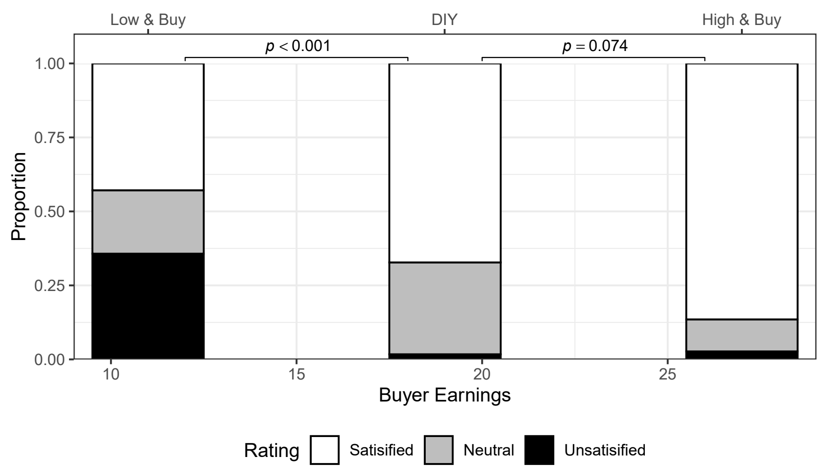
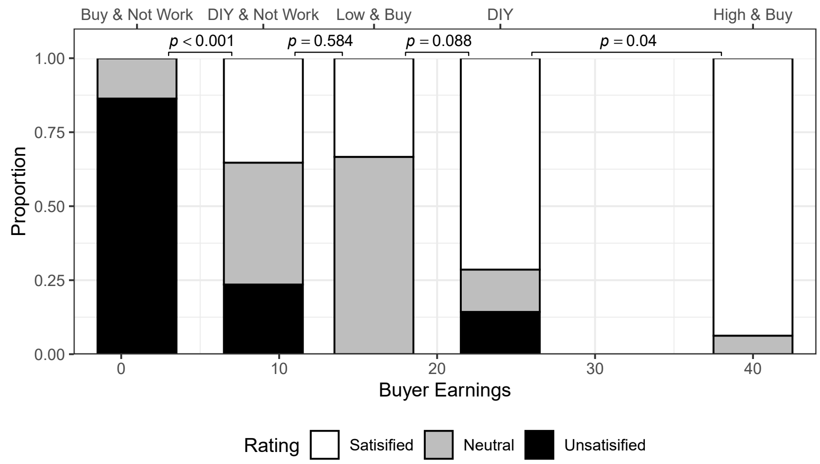


Figure 8 displays ratings given by buyers following the experts' advice in the Service Uncertainty Treatment. Each bar represents a particular level of earnings (Bottom Axis) and shows the proportions of each of the three ratings within that level. Advice, not working, and type (if the outcome reveals it) are on the Top Axis. 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 similar punishment when DIY does not work. Surprisingly, we do not see evidence of punishment when low-need buyers bought the service, i.e., when there was intentional overprovision. Additionally, there are more Unsatisfied ratings when buyers choose DIY. Brackets with p-values between bars indicate χ^2 tests of ratings across outcomes. Chi values and further tests are reported in Table A.12. Table A.11 reports counts.

Table 1 reports estimated marginal effects from an ordered probit regression on the ratings given by buyers who follow experts' advice. Table A.13 reports the coefficient estimates and cut points. Standard errors are clustered on buyers. The reference cell is the Certainty Treatment with truthful advice (not overprovision). The first row shows, in the Certainty Treatment, overprovision is penalized; relative to the

Figure 8: Rating by Outcome in Service Uncertainty Treatment



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 in 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.14 reports marginal effect from a model without the *Service Fails* variable.

Table 1: Estimated Marginal Effects from an Ordered Probit Regression on Rating

	Unsatisfied ☹	Neutral 😐	Satisfied ☺
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
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Consistent with our expectations, 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. Consistent with our expectations, 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 2 reports rates of consulting by treatment and for the reputation treatments rating. Consulting occured more frequently when the expert had a higher rating. However, some buyers consulted poorly-rated experts (the Unsatisified Column is not 0) and some buyers did not consult well-rated experts (the Satisified Column is not 1).

Table 2: Rates of Consulting by Treatment and Rating

	No Reputation	Unsatisfied ☹	Neutral 😐	Satisfied ☺
Certainty	0.731	0.400	0.500	0.821
Diagnostic Uncertainty	0.676	0.077	0.571	0.820
Service Uncertainty	0.685	0.263	0.593	0.860

Table 3 reports estimated marginal effects from a probit regression on the 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 Service Uncertainty & No Reputation are 10% less likely to consult relative to Service Uncertainty & Reputation. This difference is not statistically significant. The result is driven by negative ratings after Buy/DIY doesn't work (see below).

Table 3: Estimated Marginal Effects from a Probit Regression of Likelihood to Consult by Treatment

Treatment:		
Certainty & Reputation	-0.0328	(0.0725)
Diagnostic Uncertainty & No Reputation	-0.0557	(0.0679)
Diagnostic Uncertainty & Reputation	-0.0789	(0.0696)
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 4 reports estimated marginal effects of how ratings and treatment impact the likelihood of consulting an expert. *Ratings Seen* are 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 the 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. Similar to results from Table 3, in Mode1s 3 and 4, Diagnostic Uncertainty & Reputation is not different than Certainty & Reputation. Here

Service Uncertainty & Reputation are also not different, indicating the decrease in consultation rates for those treatments was due to the decrease in Satisfied ratings.

Table 4: 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

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

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. These results further support the second part of Finding 1. 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. These results further support the second part of Finding 2. Contrary to our hypothesis, consultation rates do not decrease when there is 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 depreciates the value of advice) and certainty is not statistically significant. These results further support the second part of Finding 3.

Table 5 reports marginal effects of a probit regression testing if there are differences across treatments in how rating seen impacts the 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 5: 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 6 reports rates of following advice to buy the high-need service by treatment and for the reputation treatments rating. Advice is followed more frequently when the expert has a higher rating. However, some buyers followed the advice of poorly-rated experts and some buyers do not followed the advice of well-rated experts.

Table 6: Rates of Buying when Advised by Treatment and Rating

	No Reputation	1*st*	Unsatisfied ☹	Neutral 😐	Satisfied ☺
Certainty	0.724	0.773	0.500	0.615	0.757
Diagnostic Uncertainty	0.607	0.774	0.125	0.286	0.833
Service Uncertainty	0.606	0.818	0.217	0.588	0.714

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, and 0 otherwise. Model 1 only includes Round 1 (no clustering) and shows there are no statistically significant differences across treatments.[16] The results are similar to the results of χ^2 tests in Figure 4. Model 2 includes all rounds; the estimates are consistent with the previous model. These results provide even more support for the second part of Findings 1, 2 and 3. 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.

Contrary to our expectations, when there was uncertainty (of either type) and the reliability of ratings was reduced, the change in the likelihood of consulting or buying given the buyer rating seen is not statistically significant. Table 8 reports marginal effects for a probit regression testing if there are differences across treatments in how rating seen impacts the 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 saw an Unsatisfied rating in the other treatments were 50% and 40% less likely to consult, none of the differences are statistically significant.

[16] 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 No Rep	-0.0528	0.0125	
	(0.0902)	(0.0730)	
Diagnostic Uncertainty Rep	-0.0968	-0.0977	-0.113
	(0.0963)	(0.0739)	(0.0897)
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

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

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

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

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 in the Certainty and Service Uncertainty Treatments[17] and 20.625 in the Diagnostic Uncertainty Treatment.[18] If buyers were completely untrusting, i.e., never bought and did not consult after the first round, earnings would be 19.67.[19] Contrary to our expectations, differences in buyer earnings between the Reputation Treatment (which theoretically increases the cost of expert dishonesty and increases buyer trust) and the No Reputation Treatment are not statistically significant. Consistent with our expectations, 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. Contrary to our expectations, buyers' ECUs are not lower (statistically significant) with service uncertainty.

Figure 9: Mean Buyer ECU by Treatment

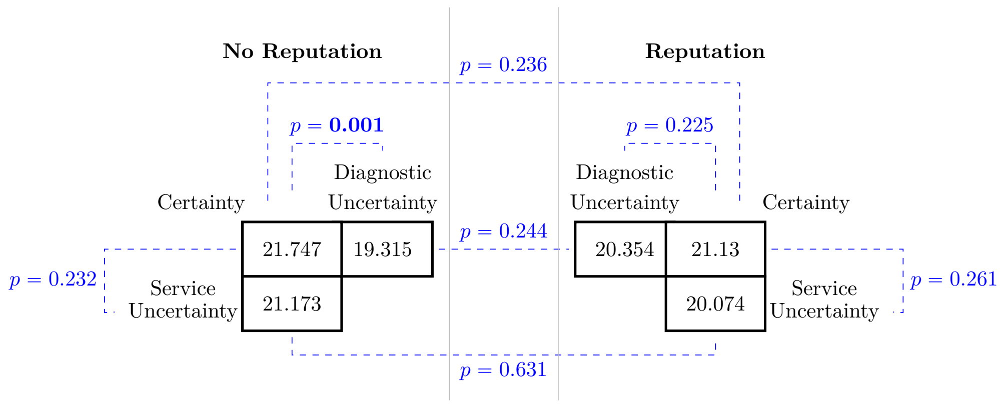


Figure 10 displays mean expert earnings in ECUs per round by treatment. Dashed lines show Wilcoxon signed-rank tests and their p-values. While none of the cross-cell tests are statistically significant, the overall (all cells pooled) test of reputation is (p-value = 0.005). For comparison, if experts were completely trustworthy and buyers were completely trusting, experts' ECUs would be 2.[20] If experts always overprovided and buyers were completely naive, i.e., 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.

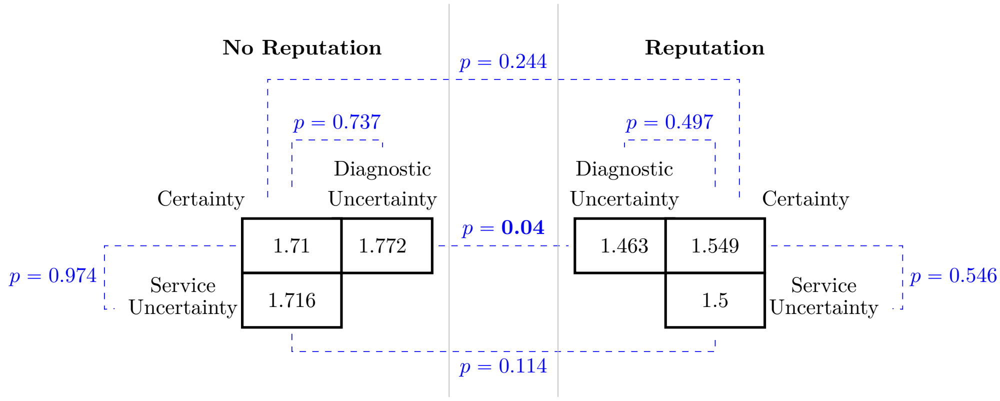
[17] Half the buyers would be low type and earn 19 ECUs and the other half would be high and earn 27. In service uncertainty, a sixth earns 9 another sixth earns 1, a third earns 24 and the final third earns 40.

[18] 1/2 would earn 19 (3/8 who would be low-need and get correct test results and advice and 1/8 who would be high-need but get inaccurate results), 1/8 would earn 11 and 3/8 would earn 26

[19] They would earn 19 in the first round, and 20 (in expectation) thereafter.

[20] 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.

Figure 10: Mean Expert ECUs by Treatment



5 Discussion

We begin our discussion with a comparison of player's behavior to best responses. Optimal behavior depends on assumptions regarding the actions of other players. Recall in Appendix A.2 we show that if a negative reputation has the maximum cost, i.e., buyers always consulted and followed the advice of positively rated experts and never consulted negatively rated experts, overprovision would not be profitable in expectation when there was certainty. However, overprovision was profitable if negative reputation was less costly. Tables 2 and 6 report the relative rates, and show the difference between positive and negative rates was not that extreme, so the cost was not high enough to make overprovision unprofitable. Despite that we see relatively low overprovision in all treatments and find no relationship between the profitability of overprovision and rates.

Turning to what is optimal behavior for the buyers, in the Certainty and Service Uncertainty Treatments, consulting has a positive expected value unless overprovision (OP) exceeds 0.75. When there is diagnostic uncertainty, the expected value is positive if the majority of experts do not overprovide ($OP < 0.5$) (see Appendix A.3). In all treatments, buying when advised has a positive expected value so long as less than 87% of experts overprovide (see Appendix A.4). Given our relatively low rate of overprovision, assuming buyers have reasonably accurate beliefs, they maximize their expected payout by trusting the expert—knowing dishonesty is rare enough that trusting is the better strategy. The majority of buyers in this experiment consulted.

The fact that the results only partially support our hypotheses highlights the importance of research in this area to help the field understand how uncertainty and reputation impact these markets. In the remainder of the discussion we review the results, particularly those that were not supported, and discuss how theory might be supplemented or amended to account for behavioral considerations suggested by the results. While we find that experts in the Reputation Treatment were as expected, more honest, we did not observe a parallel increase in buyers' trust. In the first round buy decision, buyers followed advice at higher rates in the No Reputation Treatment (than in the Reputation Treatment). This is surprising because, though the buyers in the Reputation Treatment 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 (and the experts were aware of the possibility of punishment). Even in later rounds, advice to buy was followed at lower rates in the Reputation Treatment. This rules out the possibility that buyers were cautious in the first round, but more trusting after ratings were available. One possible explanation is that the presence of a rating system alerted buyers to the possibility of expert dishonesty and absent the system they were less cautious because the possibility of dishonesty was less salient. Our results regarding reputation are consistent with those of Dulleck et al. (2011) and Schneider (2012). We, like Deck and Tracy (2020), find that buyers may rely 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.

Uncertainty inevitably makes ratings more noisy; however, the ratings have even greater noise and less reliability than we anticipated due to behavioral patterns among the buyers. With diagnostic uncertainty, when the service failed to work, buyers seemed 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 (2022) 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. However, when the buyers chose DIY and it failed, they did not blame the expert. This finding is consistent with John et al.'s (2019) finding that subjects "shoot the messenger" and punish the bearers of bad news for reporting bad outcomes that were determined independent of the messenger. 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 decisions hours before buyers via strategy method, and whatever the buyer chose the software simply implements the decisions and reports 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. In contrast, when buyers followed advice to buy the high-need service they did not actually need, but the service worked, nearly 40% of them gave the expert a Satisfactory ☺ rating. This is consistent with Mukherjee and Tracy (2022), who find that principals adjust agent bonuses based on the realization of prospects rather than norms regarding prospect selection. Despite ratings in the Uncertainty Treatments being noisier and less informative, they seem to have no less impact on the next buyer's consult and buy decisions. Our findings are also consistent with Balafoutas et al. (2020), who find that diagnostic uncertainty reduces efficiency.

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. 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). Throughout the analyses, we find a slightly lower (not statistically significant) consultation rate with diagnostic uncertainty, which is consistent with the lower value of the information.

Our results imply that social norms and concern for a positive self-image might discipline credence-goods experts even when uncertainty obfuscates deception. Healthcare, involving medical knowledge and solving rare and serious problems for clients, is the most common type of expert service observed across traditional societies and may be the most ancient form of expert service to have arisen among humans, suggesting a deep natural history of markets for healers that may have shaped our psychology for valuing mutually beneficial expert-client relationships Lightner et al. (2021, 2023); Hagen et al. (2023). Arrow (1963) noted that a unique feature of some credence goods markets, particularly those based on medical care relationships, is that consumers generally have high trust in the reliability of physicians' expert diagnoses and advice, while showing relatively little concern about questions of diagnostic uncertainty that consumers lack expert-level understanding. Expert reputation may play a role here, reflecting diagnostic reliability over time. He observes "there is an element of trust in the relation. But the ethically understood restrictions on the activities of a physician are much more severe than on those of, say, a barber (p.949)."

Elsewhere, experiments manipulating payoff uncertainty in cooperation dilemmas show that cooperation increases where there is greater payoff uncertainty van den Berg et al. (2021); Hajikhameneh and Iannaccone (2023). These observations suggest that in credence-good markets, needy consumers, unsure of their need levels and what is best for them, are more driven to consult experts and trust their advice about what is best, especially where there is greater uncertainty about whether the best treatment for them will be successful. These results indicate many people have a "preference for truth-telling" (Abeler et al., 2019).

It is possible that because uncertainty compounds buyers' confusion around the decision, it might magnify the buyer's sense of need for information and solutions (Mishel, 1988), driving expert consultation and trust in expert advice, which offsets any discounting. However, our experiment was not designed to detect or isolate the effect of the potentially competing mechanisms. 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.

Our finding that buyers are more likely to accept advice to buy when there is no rating system bears further investigation. One possible explanation is that the existence of a rating system makes the possibility of dishonesty more salient to buyers. Investigation of this possibility, beyond confirming the effect, should also investigate the impact on efficiency. Buyer naivete allows dishonest sellers to waste resources; however, overly skeptical buyers forego beneficial exchanges. Our finding that buyers' ratings are driven by the stochasticity in outcomes suggest several directions future experiments might pursue. Identifying an alternative way to elicit or frame the elicitation of ratings, so that ratings reflected an expert's propensity for honesty rather buyers' realized earnings is an important area to explore, because such a rating system could greatly improve efficiency when there is uncertainty in credence-goods markets. Another important expansion would be to introduce another rating system, particularly one that allowed averaging of ratings, e.g., stars, and expand the number of rounds. Although many individual ratings do not provide helpful information and may provide misinformation, it is possible that by aggregating ratings, the system would provide useful guidance.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond, “Preferences for Truth-Telling,” *Econometrica*, 2019, *87* (4), 1115–1153.

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

Andreoni, James, “Impure altruism and donations to public goods: A theory of warm-glow giving,” *The economic journal*, 1990, *100* (401), 464–477.

Angerer, Silvia, Daniela Glätzle-Rützle, Thomas Rittmannsberger, and Christian Waibel, “The Value of Rating Systems in Healthcare Credence Goods Markets,” SSRN Scholarly Paper ID 3965318, Rochester, NY November 2021.

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.

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. Dustin 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.

Fehr, Ernst and Klaus M Schmidt, "A theory of fairness, competition, and cooperation," *The quarterly journal of economics*, 1999, *114* (3), 817–868.

Filippas, Apostolos, John J Horton, and Joseph M Golden, "Reputation Inflation," *Marketing Science*, 2022, *41* (4), 733–745.

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.

Hagen, Edward H., Aaron D. Blackwell, Aaron D. Lightner, and Roger J. Sullivan, "*Homo medicus* : The transition to meat eating increased pathogen pressure and the use of pharmacological plants in *Homo*," *American Journal of Biological Anthropology*, April 2023, *180* (4), 589–617.

Hajikhameneh, Aidin and Laurence R. Iannaccone, "God games: An experimental study of uncertainty, superstition, and cooperation," *Games and Economic Behavior*, May 2023, *139*, 88–116.

Houser, Daniel and John Wooders, "Reputation in Auctions: Theory, and Evidence from eBay," *Journal of Economics & Management Strategy*, 2006, *15* (2), 353–369.

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

Huesmann, Katharina, Christian Waibel, and Daniel Wiesen, "Rankings in Healthcare Organizations," SSRN Scholarly Paper ID 3690851, Rochester, NY September 2020.

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.

— , Daniel Neururer, and Matthias Sutter, "Insurance Coverage of Customers Induces Dishonesty of Sellers in Markets for Credence Goods," *Proceedings of the National Academy of Sciences*, July 2016, *113* (27), 7454–7458.

— , — , and — , "Credence goods markets and the informational value of new media: A natural field experiment," *Discussion Papers of the Max Planck Institute for Research on Collective Goods Bonn 2019/3*, 2019, p. 31.

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

Lightner, Aaron D., Cynthiann Heckelsmiller, and Edward H. Hagen, "Ethnoscientific expertise and knowledge specialisation in 55 traditional cultures," *Evolutionary Human Sciences*, 2021, *3*, e37.

_ , _ , and _ , "Ethnomedical Specialists and their Supernatural Theories of Disease," *Review of Philosophy and Psychology*, June 2023, *14* (2), 611–646.

Liu, Fang, Alexander Rasch, Marco Schwarz, and Christian Waibel, "The Role of Diagnostic Ability in Markets for Expert Services," SSRN Scholarly Paper ID 3491607, Rochester, NY 2019.

Livingston, Jeffrey A., "How Valuable Is a Good Reputation? A Sample Selection Model of Internet Auctions," *Review of Economics and Statistics*, August 2005, *87* (3), 453–465.

Luca, Michael, "Reviews, Reputation, and Revenue: The Case of Yelp.com," NOM Unit Working Paper 12-016, Harvard Business School 2016.

Lucking-Reiley, David, Doug Bryan, Naghi Prasad, and Daniel Reeves, "Pennies from eBay: the Determinants of Price in Online Auctions," *Journal of Industrial Economics*, June 2007, *55* (2), 223–233.

Martinsson, Peter and Emil Persson, "Physician behavior and conditional altruism: the effects of payment system and uncertain health benefit," *Theory and Decision*, October 2019, *87* (3), 365–387.

Mimra, Wanda, Alexander Rasch, and Christian Waibel, "Price competition and reputation in credence goods markets: Experimental evidence," *Games and Economic Behavior*, 2016, *100*, 337–352. Publisher: Elsevier.

Mishel, Merle H., "Uncertainty in Illness," *Image: the Journal of Nursing Scholarship*, 1988, *20* (4), 225–232.

Mukherjee, Prithvijit and J. Dustin Tracy, "Agency, Benevolence and Justice," ESI Working Paper 22-03, Chapaman University February 2022.

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.

van den Berg, Pieter, Siegfried Dewitte, and Tom Wenseleers, "Uncertainty causes humans to use social heuristics and to cooperate more: An experiment among Belgian university students," *Evolution and Human Behavior*, May 2021, *42* (3), 223–229.

von Neumann, J, "Theory of games and economic behavior/J. von Neumann, O. Morgenstern," 1944.

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 Additional Propositions

These propositions follow from our model but not tested in this paper.

> **Proposition 4:** An increase in the proportion of high need buyers has an ambiguous effect.

Proof

$$\frac{\partial^2 \mathbb{E}[R_{i,t}]}{\partial OP_i} = (1-\alpha)\left[-\theta(1-p)sq + \phi(1-p)sq\right]$$

$$\frac{\partial^2 \mathbb{E}[R_{i,t}]}{\partial p \partial OP_i} = (1-\alpha)\left[\theta sq - \phi sq\right]$$

$$\frac{\partial \Delta \mathbb{E}[\pi]}{\partial p} = -B\mu(R_t) - \mathbb{E}_t\left[\sum_{\tau=t+1}^{T} C\lambda(\Delta R_\tau)B\mu(\Delta R_\tau)\right] \tag{5}$$

The first (immediate) term of Equation 5 is negative; there are fewer low-need buyers to overprovide to, so it is less profitable. The bracketed expected term is positive, less overprovision results in less damage to reputation and less impact on future profit. In the generic situation, it is impossible to say which will dominate.

> **Proposition 5:** An increase in the rounds of interaction, T, will not increase the profit from overprovision.

Proof (evaluate Equation 1 with an additional round)

$$(1-p)B\mu(R_t)+\mathbb{E}_t\left[\sum_{\tau=t+1}^{T}\lambda(\Delta R_\tau)\Big(C+(1-p)B\mu(\Delta R_\tau)\Big)\right] \lesseqgtr (1-p)B\mu(R_t)+\mathbb{E}_t\left[\sum_{\tau=t+1}^{T+1}\lambda(\Delta R_\tau)\Big(C+(1-p)B\mu(\Delta R_\tau)\Big)\right]$$

$$\mathbb{E}_t\left[\sum_{\tau=t+1}^{T}\lambda(\Delta R_\tau)\Big(C+(1-p)B\mu(\Delta R_\tau)\Big)\right] \lesseqgtr \mathbb{E}_t\left[\sum_{\tau=t+1}^{T+1}\lambda(\Delta R_\tau)\Big(C+(1-p)B\mu(\Delta R_\tau)\Big)\right]$$

Equation 2 does not depend on τ max, all ΔR_τ, $\tau \leqslant T$ are equal and cancel:

$$0 \lesseqgtr -\mathbb{E}_t\left[\lambda(\Delta R_{T+1})\Big(C+(1-p)B\mu(\Delta R_{T+1})\Big)\right]$$

The bracketed term is non-negative because $C > 0$, $B > 0$ and p, λ and μ are probabilities, so:

$$0 \geqslant -\mathbb{E}_t\left[\lambda(\Delta R_{T+1})\Big(C+(1-p)B\mu(\Delta R_{T+1})\Big)\right]$$

A.2 Profitability of Overprovision

This subsection calculates the expected profit from overprovision relative to honest advice, for our parameters. As in the main text, 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.2.1 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 reasons buyers following expert advice would not get the best outcome possible given their type.

Extension 1a Given $C = 1$, $B = 3$ and $T = 3$ rounds, Proposition 1 can extended to calculate the net profit from overprovision.

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

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 $0.125\times$ 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.[21] The final row is the sum (expected value $\times 8$) 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
$\sum$	$(10\lambda_\ominus + 6\lambda_\oplus)C+$ $(8\mu_0 + 10\mu_\ominus + 6\mu_\oplus)B$	$16\lambda_\oplus C+$ $(4\mu_0 + 8\mu_\oplus)B$	$10(\lambda_\ominus - \lambda_\oplus)C+$ $(4\mu_0 + 10\mu_\ominus - 2\mu_\oplus)B$

Case 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 Extension 1a, and substituting these values into Equation 6 expected net profit of lying is:

[21] 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.[22]

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

Case 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.[23] Equation 6 simplifies to:

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

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

A.2.2 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.

Case 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.[24] by Lemma 1 substituting $\lambda_\oplus = \mu_\oplus = 0.66$ and $\lambda_\ominus = \mu_\ominus = 0.33$ into Equation 6 expected net profit of lying is:

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

A.2.3 Service Uncertainty

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

Extension 3a If there is service uncertainty, the expected net profit from overprovision is,

$$\mathbb{E}[\pi(Lie) - \pi(Truth)] = \mathbb{E}[\Delta] = 64/9\lambda_\ominus - 64/9\lambda_\oplus + 108/9\mu_0 + 24/9\mu_\oplus + 192/9\mu_\ominus \tag{7}$$

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

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

[23] 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.

[24] 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**)	$\mathbf{\Delta}\pi$
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
$\sum$		$(8\mu_0 + 64/9\lambda_\ominus + 80/9\lambda_\oplus)B$ $64/9(\lambda_\ominus + 80/9\lambda_\oplus)C$	$(4\mu_0 + 8\mu_\oplus)B+$ $16\mu_\oplus C$	$(36/9\mu_0 + 8/9\mu_\oplus + 64/9\mu_\ominus)B+$ $64/9(\lambda_\ominus - \lambda_\oplus)C$

Case 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 7 expected net profit of lying is:

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

A.2.4 Empirical Confirmation

Table A.3 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 for each treatment arm and buyer need level, and then taking the mean of means, ensuring equal weight to each need level in every treatment. *Lie Gain* 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 Treatment, overprovision was more profitable (in expectation) in the no Reputation Treatment. Within no reputation, dishonesty is more profitable with uncertainty, service uncertainty more so. The final column reports the proportion of experts who recommended their service to low-need buyers.

Table A.3: Expert Earnings by Treatment and Overprovision, Balanced

Rep	Service Uncertain	Diagnose Uncertain	ECU's Over	ECU's Truth	Lie Gain	Overprovision Rate
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

A.3 Utility of Consulting

A.3.1 Certainty

In the certainty treatment, the expected utility gain (relative to not) from consulting is:

$$0.5 * (1 - OP) * (19^r) + .5 * OP * (11^r) + .5 * (27^r) - 20^r$$

where OP is the overprovision rate and we assume a utility of x^r. The first term is the utility of following advice to the properly-advised, low-need buyer. The second is the utility to over-provisioned, low-need buyers. The third term is the utility to high-need buyers.

The final term is the utility of DIY, 20 ECU for certain.

Assuming risk neutrality ($r = 1$)

$$0.5 * (1 - OP) * 19 + .5 * OP * 11 + .5 * 27 - 20$$

$$(1 - OP) * 19 + OP * 11 + 27 - 40$$

$$19 - OP * 19 + OP * 11 + 27 - 67$$

$$6 - OP * 8$$

$$0.75 - OP$$

is 3 if $OP = 0$ and is only negative if OP exceeds 0.75. Risk aversion decreases the value of consulting; at $r = 0.5$, is negative if OP exceeds ~ 0.586.

A.3.2 Service Uncertainty

In the service uncertainty treatment, the expected utility gain (relative to not) from consulting is:

$$.5 * (1 - OP) * (\frac{2}{3} * 24^r + \frac{1}{3} * 9^r) + .5 * OP * (\frac{2}{3} * 16^r + \frac{1}{3} * 1^r) + .5 * (\frac{2}{3} * 40^r + \frac{1}{3} * 1^r) - (\frac{2}{3} * 25^r + \frac{1}{3} * 10^r)$$

Like certainty, if the buyer is risk-neutral consulting has a positive expected value (3) unless OP exceeds 0.75. However, when the buyer is risk averse (r=0.5), even without overprovision consulting net gain is only ~ 0.02, and only has a positive expected value when OP is below ~ 0.03225.

A.3.3 Diagnostic Uncertainty

In the service uncertainty treatment, the expected utility gain (relative not) from consulting is:

$$(1 - OP) * (0.5 * 19^r + 0.5 * s * 27^r + 0.5 * (1 - s) * 11^r) + OP * (0.5 * 11^r + 0.5 * 27^r) - 20^r$$

where s is the probability the signal is correct. Given $s = 0.75$, the value of consulting is only 1 ECU. It remains positive if the majority of experts do not overprovision ($OP < 0.5$). Risk aversion ($r = 0.5$) reduces the expected value of consulting to ~ 0.31. It has a positive value when OP is below ~ 0.246.

A.4 Utility of Buying

A.4.1 Certainty

In the Certainty and Service Uncertainty Treatments, the expected utility gain of buying (relative to not) when advised to is:

$$0.5 * [(1 - OP) * 19^r + OP * 11^r] + .5 * 27^r - 19^r$$

Assuming risk neutrality ($r = 1$)

$$0.5 * [(1 - OP) * 19 + OP * 11] + .5 * 27 - 19$$

$$0.5 * [19 - 19 * OP + 11 * OP] + .5 * 27 - 19$$

$$0.5 * [19 - 8 * OP] + .5 * 27 - 19$$

$$4 - 8 * OP$$

If the buyer is risk neutral, buying when advised to has positive expected value provided OP < 0.5

A.4.2 Diagnostic Uncertainty

In the Diagnostic Uncertainty Treatment, the expected utility gain of buying relative to not when advised to is:

$$0.5 * [(1 - OP) * 19^r + OP * (0.75 * 11^r + 0.25 * 27^r)] + 0.5 * 0.75 * 27^r + 0.5 * 0.25 * 11^r - 19^r$$

Assuming risk neutrality ($r = 1$)

$$0.5 * [(1 - OP) * 19 + OP * (0.75 * 11 + 0.25 * 27)] + 0.5 * 0.75 * 27 + 0.5 * 0.25 * 11 - 19$$

$$0.5 * [19 - OP * 19 + OP * 15] + 11.5 - 19$$

$$0.5 * [19 - OP * 4] + 11.5 - 19$$

$$2 - 2 * OP$$

The risk-neutral buyer has a positive expected value in taking advice to buy unless there is complete overprovision and then the expected value is 0.

A.5 Participant Characteristics

Table A.4: Subject Characteristics

Age Range	18-25	26-45	46-64	65+	
Obs.	241	311	65	7	
Sex	Female	Male	Other		
Obs.	357	250	17		
Education	$\leqslant$ HS	Some College	Associates	Bachelor	Grad School
Obs.	82	177	53	206	106

A.6 Instructions within Software

There were separate instructions for experts and buyers. This was done to aid participant comprehension. Both instructions explain the complete game form.

A.6.1 Buyer Instructions

Your role is **BUYER**.

Your tasks are:

1. decide whether or not to consult an expert
2. decide whether to buy a service from them or "do-it-yourself" (DIY)

In each of the **three rounds** you participate in, you are randomly assigned a need type (Low or High with equal probability) which you do not know.

[The service and the DIY solution only work 2 out of 3 times. | Treatment == Service Uncertainty] Based on the payment tables below,

- if you are **low need** type, you get the same benefit [+10 ECU | Treatment != Service Uncertainty] [+15 ECU (if it works)| Treatment == Service Uncertainty] from buying the service or DIY, so you **are better off with the cheaper DIY option**
- if you are **high need** type, you get greater benefit from buying the service [+26 ECU | Treatment != Service Uncertainty] [+39 ECU (if it works)| Treatment == Service Uncertainty] than from DIY [+10 ECU | Treatment != Service Uncertainty] [+15 ECU (if it works)| Treatment == Service Uncertainty], so you **are better off buying the service**

In the first round, you consult a randomly selected Expert who tests your need type. **The test is** [**always right** | Treatment != Diagnostic Uncertainty] [**right 3 out of 4 times** | Treatment == Diagnostic Uncertainty]. Based on your test result, the Expert advises you whether to buy a service or DIY (the advice is the same for all Buyers with the same test results). In the second and third round you can choose whether or not to consult an Expert you are randomly paired with in that round.

Each round, you start with 10 ECU. Consulting the Expert costs you 1 ECU and buying the service costs you an additional 8 ECU. The Expert earns more if you consult them and buy the service. They can choose to advice either buying the service or DIY, regardless of the test results.

[At the end of each round you will rate the Expert. In the second and third rounds you will see the rating from the last Buyer who consulted that Expert before you make decisions. | Treatment == Reputation]

The table below summarizes costs and benefits based on Buyer's need and decisions. [Note that the service and the DIY solution only work 2 out of 3 times.| Treatment == Service Uncertainty]

< Either Figure 1 or Figure 2 >

You will get paid $1.30 for successful completion of the study plus any additional earnings from your interactions with the Experts.

The expected payment is about the same for Buyers and Experts. For Buyers, **1 ECU equals $0.04.**

A.6.2 Expert Instructions

Your role is **EXPERT**.

You will be matched with three consecutive Buyers. For each, you will earn:

- 0 ECU if they do not consult you
- 1 ECU if they do consult you
- 9 ECU if they consult you and buy your service

Half the Buyers are **low need** type, and **half** are **high need** type (randomly determined). They do not know their need type. If they consult you, you get a test result with their type. **The test is** [**always right** | Treatment == Diagnostic Uncertainty] [**right 3 out of 4 times** | Treatment != Diagnostic Uncertainty].

Your task is to pick advice for both low and high need type Buyers. This advice will be given to all Buyers with the same test results.

- **Low** need type get as much benefit from a cheaper "do-it-yourself" (DIY) solution as from your service, so [on average | Treatment == Diagnostic Uncertainty] they **are better off choosing DIY**.
- **High** need type benefit more from your service than from the DIY, so [on average | Treatment == Diagnostic Uncertainty] they **are better off buying the service**.

The first Buyer will consult you.

[All Buyers who consult you rate their satisfaction with you:

☺	😐	☹
Satisfied	Neutral	Unsatisfied

| Treatment == Reputation]

The second and third Buyers [will see the last available rating that a buyer gave you. They | Treatment == Reputation] have the option of not consulting you.

You will be paid $1.30 for successful completion of the study. Other participants in this study will play the role of Buyers matched to you and after they complete their task you will receive any additional earnings, along with your payment for survey completion, within 1 week.

The expected payment is about the same for Buyers and Expert. For Experts, **1 ECU equals $0.90**

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. | Treatment == Service Uncertainty]

< Either Figure 1 or Figure 2 >

A.6.3 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.7 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.8 Ratings

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

Table A.5: Ratings Given by Treatment

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

Table A.6 reports χ^2 test for equal distributions of ratings across all treatment combinations.

Table A.6: 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.7 reports ratings by the ECU's the buyer earned through their decisions for buyers following the experts' advice in the Certainty Treatment.

Table A.7: 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.8 reports χ^2 test for equal distributions of ratings across all earning combinations within the Certainty Treatment.

Table A.8: 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.9 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.9: 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.10 reports χ^2 test for equal distributions of ratings across all earning combinations within the Diagnostic Uncertainty Treatment.

Table A.10: 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.11 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.11: 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.12 reports χ^2 test for equal distributions of ratings across all earning combinations within the Service Uncertainty Treatment.

Table A.12: 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.13 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 (coefficient estimates) when over-provision is interacted with an indicator for Diagnostic Uncertainty Treatment and none with Service Uncertainty Treatment, 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 buyers penalize the experts when the selection does not work, despite that this was random and independent of the experts. This explains the negative rating for truthfulness in service uncertainty. This confirms graphically what Figure 8 shows.

Table A.13: 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.14 reports marginal effect from a model without the *Service Fails* variable.

Table A.14: Estimated Marginal Effect from Ordered Probit Regression on Rating

	Unsatisfied ☹	Neutral 😐	Satisfied ☺
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

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