Angerer, Silvia; Glätzle-Rützler, Daniela; Mimra, Wanda; Rittmannsberger, Thomas; Waibel, Christian

Working Paper

The value of rating systems in credence goods markets

Working Papers in Economics and Statistics, No. 2025-03

Provided in Cooperation with:
Institute of Public Finance, University of Innsbruck

universität innsbruck

Faculty of
Economics and Statistics

The Value of Rating Systems in Credence Goods Markets

Silvia Angerer, Daniela Glätzle-Rützler, Wanda Mimra, Thomas Rittmannsberger, and Christian Waibel

Working Papers in Economics and Statistics

2025-03

University of Innsbruck
https://www.uibk.ac.at/EconStat/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
Faculty of Economics and Statistics
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 96136
E-mail: Dean-EconStat@uibk.ac.at

The most recent version of all working papers can be downloaded at
https://www.uibk.ac.at/fakultaeten/volkswirtschaft_und_statistik/forschung/wopec/

For a list of recent papers see the backpages of this paper.

The Value of Rating Systems in Credence Goods Markets*

Silvia Angerer, Daniela Glätzle-Rützler, Wanda Mimra, Thomas Rittmannsberger, and Christian Waibel†

Abstract

In this paper, we experimentally investigate the effect of public consumer ratings on market outcomes in credence goods markets. Contrary to search or experience goods, consumers cannot evaluate all dimensions of trade for credence goods, which may inhibit the information and reputation-building value of public rating systems. We implement a market in which experts have an informational advantage over consumers with respect to the appropriate service level. The rating system takes the form of a five-star rating system as is common on online rating websites. The value of this rating system is compared in two different expert market settings: First, one in which consumers cannot rely on information from personal experience with the expert, reflecting markets in which consumer-expert interactions are often first-time and infrequent (e.g. specialist visits in healthcare markets). Second, one in which consumers have personal experience with the expert, reflecting markets in which consumer-expert interactions are frequent and repeated (e.g. general practitioner visits in healthcare markets). We find that the public rating system significantly improves market outcomes. Furthermore, a public rating system is a good substitute for personal experience information in terms of market efficiency and consumer surplus. Combined, however, we find no complementarity between public ratings and personal experience information, mainly due to the already high market efficiency in the presence of either one.

Keywords: Credence goods, expert behavior, ratings, feedback, laboratory experiment
JEL classification: C91, D82, I11, L15

* We thank seminar and conference participants at the 7th Workshop in Behavioral and Experimental Health Economics in Innsbruck, the European Health Economics Workshop 2024 in Toulouse, the 5th ATHEA Conference in Vienna, the SABE 2020 Annual Conference, the EuHEA 2020 Seminar Series, the dggö Workshop, the gesundheitsökonomische Ausschuss of the Verein fuer Socialpolitik, in Nancy, Groningen and Paris for their helpful comments. Especially, we want to thank Geir Godager, Nadja Kairies-Schwarz, Mathias Kifmann and Robert Nuscheler for their discussion and input. Financial support from the Nachwuchsförderung at the University of Innsbruck as well as from the Austrian Central Bank (Jubilaeumsfonds Project 17805) is gratefully acknowledged. The project was approved by the internal review board of the University of Innsbruck.

† Angerer: UMIT TIROL - Private University for Health Sciences and Health Technology, *silvia.angerer@umit-tirol.at*; Glätzle-Rützler: University of Innsbruck,*daniela.ruetzler@uibk.ac.at*; Mimra: ESCP Business School,*wmimra@escp.eu*; Rittmannsberger: Technical University of Munich, *thomas.rittmannsberger@tum.de*; Waibel: ETH Zürich, *cwaibel@ethz.ch*

1. Introduction

In many markets, customers rely heavily on expert advice as they lack the ability to assess the quality they require. Healthcare serves as a prime example: While physicians are experts concerning the appropriate quality of service, patients typically do not know which treatment they need. Often, patients cannot even verify the adequacy of the provided treatment ex-post.

Services (or goods) involving an asymmetric information problem between an expert seller and uninformed customers regarding adequate quality are referred to as "*credence goods*" (Darby & Karni, 1973; Dulleck & Kerschbamer, 2006). In these markets—such as financial services, repair services, legal advice, and healthcare services—significant inefficiencies may arise: Depending on market institutions and (financial) incentives, experts may *overtreat* by providing unnecessary services, or *undertreat* by providing insufficient services. Furthermore, experts may *overcharge* by billing for services that were not provided.[1] Concerns about expert behavior are particularly strong in markets for healthcare services, due to their societal and economic significance at roughly 10% of GDP on average in OECD countries (OECD, 2023).[2]

As asymmetric information between expert sellers and their customers is the key friction in credence goods markets, providing information to customers can potentially alleviate market inefficiencies.[3] However, this depends on the precise nature of information, in light of the fundamental problem that certain dimensions of expert and service quality cannot be judged even after consumption. This paper analyzes the effects of an important and increasingly prominent form of information in credence goods markets: a public rating system of experts.

Feedback platforms like Yelp, Google, TripAdvisor, Uber, etc., where consumers can rate their experiences with an expert, have become more and more prominent in recent years. Yelp, for instance, counts approximately 28 million monthly users and has accumulated over 214 million customer reviews since its introduction in 2004, nine percent of them in the area of healthcare (Yelp, 2020). These platforms provide consumers with relevant information when choosing experts, such as physicians (Xu et al., 2021). Given the widespread utilization of rating websites and the spare empirical evidence on their effectiveness in improving market outcomes, studying this particular form of information—previous consumers' feedback in the form of an expert rating—is of importance in credence goods markets.

[1] Balafoutas & Kerschbamer (2020) provide a comprehensive review of the recent literature on credence goods.

[2] According to Brown & Clement (2018), a sizable part of healthcare expenditures may be unnecessary. Brown & Clement (2018) categorize 1.52 million healthcare services administered between July 2015 and June 2016 in Washington state into 3 categories (necessary, likely wasteful, and wasteful) and conclude that 44% of those are deemed wasteful, amounting to excess spending of $258 million (33% of the total $785 million spent on health care services).

[3] In early work, Domenighetti et al. (1993) investigate the information channel in the healthcare sector by analyzing the treatments for better-informed customers, i.e. patients that are physicians.

Public rating information is however not the only information available to a consumer before deciding to visit an expert. The consumer may have consulted the expert before and thus have some previous experience with a particular expert. For instance, in healthcare markets, patients typically are in repeated interaction with general practitioners. On the other hand, numerous expert visits occur rarely or only once, such as specialist consultations, which means that consumers may not be personally acquainted with the expert and may solely rely on publicly available information, if any.[4] In this paper, we analyze the value of a public rating system of experts in these two different market environments, when consumers have access to personal experience information with an expert, and when this is not the case. In particular, the experimental design allows for comparing these different types of information as well as analyzing their interaction.

We implement a credence goods laboratory experiment with two frames: a healthcare market frame and a neutral frame. Experts and consumers interact over 16 periods in a classic credence goods market set-up in which experts have short-term incentives to undertreat and overcharge.[5] In particular, a consumer has a problem that needs to be treated but does not know the severity of it. Experts can costlessly diagnose the problem and provide and charge for either a minor or a major treatment. In this setting, information about past expert behavior opens up the possibility of reputation-based equilibria. The focus of this paper is in particular how information in the form of a public rating system provides these reputation incentives, on a stand-alone basis and in comparison to personal experience information.

To keep the set-up simple and focus on the reputational effects of ratings, we fixed the prices and therefore mark-ups in the experiment.[6] The public rating system is implemented as the consumer's choice to give feedback on a zero to five-star scale. In particular, after having received treatment from an expert and being charged a price, consumers observe their payoff and can decide to provide a rating. These ratings are then averaged and provided to consumers before their decision to visit an expert in the next period in the rating conditions.[7] To distinguish between markets with and without personal experience information, experts can be either identified by a fixed ID (personal experience conditions) or not. This determines whether personal experience—payoffs from previous interactions—can be attributed to given experts and thus used to select and incentivize particular experts.

[4] In healthcare markets, the degree to which a physician has more repeated interactions compared to first-time or one-shot interactions depends among others on the specialty. Physicians performing rare examinations (e.g. radiologists doing MRI or CT scans) will have more first-time or one-shot interactions compared to general practitioners for instance.

[5] Experts could also overtreat, but given the parametrization, this is dominated by simply overcharging instead of overtreating.

[6] This shuts down the potentially confounding effects of price competition. Mimra et al. (2016) show that competition for prices undermines reputation-building incentives for experts in credence goods markets. Fixed prices are also in line with the fact that prices are heavily regulated in credence goods markets, notably healthcare markets.

[7] Throughout, we use the term condition for experimental treatments so as to not create confusion with the standard credence goods terminology of a treatment given by the expert to solve the consumer's problem.

We find that a public rating system significantly improves efficiency and consumer surplus, independent of the market framing (i.e. neutral or health care frame): Compared to a baseline in which neither a public rating system nor personal experience is available, both undertreatment and overcharging decrease significantly. The latter result is particularly interesting, as contrary to undertreatment, overcharging cannot be detected by consumers. Furthermore, we find that a public rating system is a good substitute for personal experience information: Market efficiency and consumer surplus are on the same levels in markets with a public rating system compared to personal experience markets. Thus, in expert markets that are characterized by many first-time or infrequent interactions in which consumers cannot rely on their own past experience to choose experts, a public rating system proves to be a well-functioning information alternative even in credence goods markets. Finally, we do not find complementarity between public rating and personal experience information when combined: Market outcomes do not improve further. However, this might be due to the fact that efficiency is already at a very high level when either type of information is available to choose and incentivize experts.

Our main contribution is to provide causal evidence on the effectiveness of a public rating system in credence goods markets. To the best of our knowledge, there exists no study systematically investigating the effect of public rating systems on expert behavior in a credence goods setting, and no study that disentangles the effect of public rating systems for the two different information environments. Recent research on experience goods suggests that, while public rating systems are beneficial in the first situation (Tadelis, 2016), they do not carry many additional benefits when market participants draw on personal relationships (Cai et al., 2014). Little is known, however, on the effectiveness of public rating systems in credence goods markets, either in general or in a healthcare setting in particular.

Our motivation to take the problem to the lab is essentially twofold: An experiment allows to design and fully control for the asymmetric information problem as well as to implement different market institutions. Thus, in contrast to observational data, a controlled laboratory experiment provides the advantage of observing the consumer's "true" problem and therefore unambiguously classifying expert behavior. Even though the setting does not take into account all factors of an expert-consumer relationship, the laboratory offers a clean testbed for introducing different institutions and disentangling reputational incentives, which is difficult using observational data as these are not cleanly separated. Additionally, an experiment allows to measure the effect of introducing a public rating system on market outcomes such as efficiency and consumer surplus.

The paper proceeds as follows: The next section discusses the related literature, followed by the experimental design. In the results section, we present the combined results of the two market frames: a neutral market frame and a healthcare market frame. Section 5 concludes.

2. Related literature

Following the pioneering works on credence goods markets by Darby & Karni (1973), Dulleck & Kerschbamer (2006), and Dulleck et al. (2011), several studies set out to analyze the impact of different institutions such as competition, reputation, second opinions, price regulations, insurance coverage, new media, or monitoring. The papers conclude that several institutions could potentially mitigate inefficiencies in credence goods markets (Angerer et al., 2021; Balafoutas et al., 2013; Balafoutas & Kerschbamer, 2020; Balafoutas et al., 2017; Huck et al., 2016a; Kerschbamer et al., 2016, 2017; Liu et al., 2021; Mimra et al., 2016; Rajgopal & White, 2019). In what follows, we shortly introduce and discuss studies investigating the impact of reputation.

Building upon the seminal papers by Klein & Leffler (1981), Kreps et al. (1982), and Shapiro (1982), a large and growing body of literature has investigated the effects of direct and indirect reputation in experience goods markets (e.g., Bar-Isaac & Tadelis, 2008; Bohnet & Huck, 2004; Bolton et al., 2004; Ely et al., 2008; Ely & Välimäki, 2003; Tadelis, 2016). There are three papers on experience goods closely related to our present work by Bohnet & Huck (2004), Huck et al. (2012), and Huck et al. (2016b). Letting subjects play a binary-choice trust game for 20 periods, Bohnet & Huck (2004) find that direct reputation is more effective in promoting trust than indirect reputation. Extending their model, allowing for competition between trustees, Huck et al. (2012) conclude that competition, coupled with direct reputation, helps eliminate market misconduct completely. However, Huck et al. (2016a) show that incentives for reputation-building are diminished once trustees start competing over prices.

The key difference between trust games and markets for credence goods is that, although participants in trust games have asymmetric information ex-ante, information is symmetric ex-post, whereas credence goods markets are characterized by persistent information asymmetries. Due to this, reputation-building may be impeded in credence goods markets, as experts have no way of unambiguously signaling trustworthiness to potential customers. The notion of credence goods was first introduced by Darby & Karni (1973). In their seminal paper, Dulleck & Kerschbamer (2006) provide a unifying theoretical framework and investigate the effectiveness of different institutions in markets for credence goods, among others (direct) reputation and competition, tested experimentally in Dulleck et al. (2011) under flexible prices. They find that, while competition drives down prices, therefore benefitting customers, it does not enhance overall market efficiency as undertreatment, overtreatment, and overcharging rates do not improve, compared to a situation without competition. Neither (direct) reputation nor a combination of (direct) reputation and competition influences relevant market outcomes under flexible prices. Conducting a field experiment in the U.S. market for auto repairs, Schneider (2012) concludes that reputation does not improve market outcomes in credence goods markets. In a recent literature review, Balafoutas & Kerschbamer (2020) find that the impact of

competition and reputation on expert behavior in credence goods markets is at best ambiguous. The paper on credence goods closest to the present study is by Mimra et al. (2016). They experimentally investigate the role of reputation in markets under different price regimes (price competition and fixed prices) and with two forms of reputation mechanisms (private and public histories). Under private histories, customers receive information on posted prices, charged prices, whether undertreatment occurred, and their period payoff for their own previous interactions with an expert. Under public histories, customers receive this information for all previous interactions of an expert including their own. Note, that no environment without the possibility to build a direct reputation is studied. The authors find that, regardless of the underlying reputation mechanism, undertreatment is significantly higher in markets with price competition compared to those under fixed-price regimes. Reputation through public histories has no impact compared to private histories in either of the price regimes. They conclude that price pressure undermines reputation-building, explaining why regulating prices may increase consumer welfare in credence goods markets.

Our main contribution to the existing literature on institutions in credence goods is that we experimentally test how a public rating system of experts, where customers can rate interactions with experts on a five-star rating scale, influences outcomes under a fixed-price regime. We can thereby distinguish the effect in two relevant market settings, markets of first-time interactions without personal experience information and those in which customers have personal experience information.

More recently and following the rise in online markets (such as eBay, Amazon, etc.), there has been an increased interest in electronic reputation systems (Ba & Paul, 2002; Bolton et al., 2004; Cabral & Hortaçsu, 2010; Dellarocas, 2003, 2006; Moreno & Terwiesch, 2014; Resnick & Zeckhauser, 2002; Resnick et al., 2006; Rice, 2012). Online markets lacked traditional reputation, but electronic reputation systems were designed to enhance trust and cooperation and to facilitate the exchange of information about the quality and reliability of market participants. Consumers can provide feedback on sellers' goods/services, creating aggregated ratings that reflect the seller's past performance and allow them to build a reputation. There is a growing body of research on electronic reputation mechanisms in experience goods markets, with studies examining their effects on market outcomes such as prices (Ba & Paul, 2002; Moreno & Terwiesch, 2014; Resnick et al., 2006), trading volume (Cabral & Hortaçsu, 2010; Moreno & Terwiesch, 2014), and seller performance (Bolton et al., 2004; Rice, 2012). Some studies have shown that reputation systems can reduce information asymmetry, increase trust (Dellarocas, 2003), and increase competition among sellers (Cabral & Hortaçsu, 2010). The findings in this literature are mainly based on laboratory experiments where students play a trust game (Bolton et al., 2004; Rice, 2012), field experiments on online trading platforms such as ebay.com (Resnick & Zeckhauser, 2002; Resnick et al., 2006), analyzing observational data from such platforms (Ba & Paul, 2002; Cabral & Hortaçsu, 2010; Dellarocas, 2005). Over the past few years,

many rating platforms were introduced for *offline* markets which enable consumers to provide feedback and rate the expertise of providers across various goods and services markets. These platforms have become particularly relevant in credence goods markets, such as healthcare, repair, and legal services.[8]

While rating systems have been shown to have a positive impact on experience goods markets, it remains an open question whether ratings will be as effective in credence goods markets. This is due to the fact that consumers are unable to determine whether the quality of the product or service provided was suitable. Our main contribution to the literature on electronic reputation and feedback systems is that we expand it to credence goods markets and experimentally test the value of a public rating system of experts in two different market settings, one with, and one without personal experience.

In a recent field experiment in the computer repair market, Kerschbamer et al. (2023) analyze the effect on the repair price of consumers signaling self-diagnosis information from the internet, and also look at the correlation between online ratings and repair prices. Kerschbamer et al. (2023) find that for reliable ratings, such as those classified as recommended by Yelp, better-rated shops charge lower prices, while the opposite is observed for non-recommended ratings. With our experiment, we contribute to this literature by providing causal evidence of the effects of the introduction of rating systems on the quality of expert services.

Given that healthcare services represent an important credence goods market, and our experiment specifically applies a healthcare frame (in addition to a neutral frame), our work also contributes to the literature on financial incentives of healthcare professionals, physician decision-making, and experimental health economics. Gruber & Owings (1996), one of the first works in the field, demonstrated that healthcare providers respond to financial incentives. This assertion has been further corroborated by a mounting body of empirical evidence, indicating that physicians and other healthcare professionals react to financial incentives with potentially adverse welfare effects (Anthun et al., 2017; Baker, 2010; Barros & Braun, 2017; Batty & Ippolito, 2017; Chao & Larkin, 2022; Clemens & Gottlieb, 2014; Dafny, 2005; Dai et al., 2017; Dunn & Shapiro, 2014; Geruso & Layton, 2019; Iizuka, 2007; Januleviciute et al., 2016; Parkinson et al., 2019; Shigeoka & Fushimi, 2014). Undertreatment, for instance, has been shown in the area of pain management (Pasero & McCaffery, 2001), for the introduction of a fixed-price prospective payment system (Cutler, 1995) as well as for uninsured patients visiting a hospital after a severe car accident (Doyle, 2005). Evidence for overtreatment is provided by Gottschalk et al. (2020) in a field experiment in the dental care market, where every fourth dentist visit resulted in the recommendation of unnecessary fillings. Overcharging happens for instance through upcoding in DRG-based hospital reimbursement systems[9] (Cook & Averett, 2020; Jürges & Köberlein,

[8] See for example www.jameda.de, www.yelp.com, or www.lawyers.com.
[9] Diagnosis-related group (DRG) is a case classification system for the reimbursement of inpatient care.

2015).[10]

In experimental health economics, one main focus is the comparison of physician behavior under different payment schemes, following the seminal article by Hennig-Schmidt et al. (2011). Hennig-Schmidt et al. (2011) and several further laboratory experiments find that there is overtreatment under fee-for-service and undertreatment under capitation payment schemes (Brosig-Koch et al., 2016, 2013, 2017b,c; Green, 2014; Lagarde & Blaauw, 2017). [11] We contribute to this literature with the study of dynamic incentives via the rating system.

Lastly, the paper contributes to the evolving literature on the value and reliability of (online) rating mechanisms in healthcare markets. A considerable amount of studies looked at the association between online physician ratings and other quality measures. While some find associations between them (Lu & Rui, 2018), others don't (Saifee et al., 2019, 2020). Conducting a systematic literature review, Hong et al. (2019) conclude that the relationship between physician ratings and clinical outcomes is at best weak. Interestingly, Saifee et al. (2020) argue that they perform poorly, especially in disciplines characterized by extensive credence goods nature (e.g., chronic disease care) because there it is particularly difficult for patients to assess the effectiveness of a particular physician accurately, given the long treatment horizon.

3. Experiment

The experimental design is based on the credence goods framework of Dulleck & Kerschbamer (2006) and the seminal experiment by Dulleck et al. (2011). In the main experiment, we implemented two framings: A neutral market frame similar to Dulleck et al. (2011), as well as a healthcare market frame, representing one of the most important credence goods markets.[12] The experimental instructions with the healthcare frame referred to expert sellers as physicians and consumers as patients, and the service for which there is asymmetric information is a treatment for a health problem. Throughout the paper, we will use the wording of 'expert' on the supply side and 'consumer' on the demand side.

[10] Further field experimental support for biased expert decisions in healthcare markets is provided by Chen & Goldman (2016), Currie et al. (2014), Currie et al. (2011), Das & Hammer (2007), Das et al. (2016), and Lim et al. (2002).

[11] Other laboratory experiments in the context of health economics look at the impact of insurance (Huck et al., 2016a), performance disclosure (Godager et al., 2016), non-monetary incentives (Kairies & Krieger, 2013), professional norms (Kesternich et al., 2015), competition between healthcare providers (Brosig-Koch et al., 2017a; Han et al., 2017), and whether teams of decision-makers decide differently than individuals (Han et al., 2020). For a comprehensive review of behavioral experiments in health economics see (Galizzi & Wiesen, 2018).

[12] The healthcare market framing is applied based on the insights of Angerer et al. (2023), Kairies-Schwarz et al. (2017), Kesternich et al. (2015), and Reif et al. (2020) who explore the effect of different framings in economic laboratory experiments.

3.1. The basic setup and parametrization

In the basic setup, experts and consumers are grouped in a market of eight subjects, four consumers, and four experts. Consumers suffer from a major problem with probability $h = 0.5$ and a minor one with probability $(1-h)$. The probability h is common knowledge. Consumers decide whether to consult an expert knowing that they suffer from some problem in every period. They do not get information about the severity of their problem. Experts diagnose their consumers' problems with certainty and at zero costs. They provide one of two treatments, a major treatment (q_H) or a minor treatment (q_L). The cost for the expert to provide the major treatment is 6 ECUs.[13] The cost for the minor treatment is 2 ECUs. Treatment prices, paid by the consumers, are either 8 ECUs (p_H) or 3 ECUs (p_L) respectively. The major treatment solves both, the major and the minor problem, while the minor treatment only solves the minor one. Consumers obtain 10 ECUs (v) if their problem is solved, and zero if treated insufficiently. The payoff for consumers consulting an expert is the difference between the obtained value and the price charged (p_H or p_L). For experts, the payoff is the spread between the price charged (p_H or p_L) and the cost of the chosen treatment.[14] In case a consumer decides against consulting any expert, the consumer receives an outside option of (-4) ECUs (o_{Con}). Experts receive $o_{Exp} = 0$ if they do not interact with any consumer in a given period. Compared to the framework of Dulleck et al. (2011), our basic model differs in two dimensions. First, the outside option of consumers is negative ($o_{Con} = -4$) illustrating the disutility of an unsolved problem.[15] Second, p_H and p_L are exogenously fixed, which is common in many expert markets, notably in highly regulated healthcare markets. Throughout the experiment, there is neither verifiability nor liability, allowing us to investigate both undertreatment and overcharging.

The structure of the stage game is as follows:

1. For each consumer, nature draws the type of problem. With probability h consumers have a major problem, and with probability $(1-h)$ consumers have a minor problem.

2. Consumers decide whether to consult an expert. If consumers decide not to visit an expert, the period ends. Otherwise, they choose one expert from a list of four.[16]

3. Experts costlessly diagnose the problem, provide a treatment (q_H or q_L), and charge a price (p_H or p_L). [17]

[13] Experimental Currency Unit (ECU)

[14] Following Dulleck et al. (2011), we assume large economies of scope between diagnosis and treatment. Hence, consumers who decide to consult an expert commit to undergo treatment by this expert.

[15] This negative outside option ensures market interaction, facilitating the investigation of the effect of ratings.

[16] Depending on the experimental condition, experts can be identified through a personal ID (in the personal experience conditions) and/or the average rating from previous periods is displayed at this stage for each expert (in the rating conditions).

[17] Depending on the experimental condition, consumers can be identified through a personal ID (in the personal experience conditions) and/or the average rating from previous periods is displayed at this stage for each expert (in the rating conditions).

4. Consumers and experts observe their payoff in the respective period.

5. In the conditions with a public rating system after learning the payoff for the respective period, consumers decide whether to rate the interaction with the expert. If they decide to rate the interaction, they choose the rating on a scale between 0 and 5 stars which is shown to the expert.

The stage game is played for 16 periods in all experimental conditions.

3.2. Experimental conditions

We employ a 2×2 factorial design to test the effect of a public rating system as outlined in Table 1. The value of a public rating scheme is analyzed and compared in two different expert market environments: First, a market environment in which consumers can, over time, rely on their personal experience with a particular expert. Second, a market environment in which consumers cannot rely on their personal experience with a particular expert. The latter represents markets in which consumer-expert interactions are often first-time and infrequent (such as specialist visits in healthcare markets), whereas the former represents markets in which consumer-expert interactions are more frequent and repeated (such as general practitioner visits in healthcare markets).

Table 1: Experimental conditions

		Market environment: Personal experience with expert	
		No	Yes
Public rating	No	***Baseline***	***Experience***
	Yes	***Rating***	***Exp+Rating***

Note: In all our experimental conditions experts compete for consumers, i.e., consumers choose one expert from a list of four if they decide to visit an expert.

These two different market environments are implemented in the experiment as follows: In the experimental conditions without personal experience with experts, in each period consumers choose one expert from a list of four without being able to identify them. All players are informed beforehand that consumers have no means of identifying experts from previous periods. Thus, although consumers observe their payoffs in each period and can partially infer expert behavior, they cannot attribute it to a particular expert and therefore cannot build up personal experience with a particular expert. In the experimental conditions with personal experience, consumers can on the contrary identify experts by a fixed ID and decide whether

to interact with a particular identified expert. Over the 16 periods of play, they can thus learn from their personal experience (payoffs) with a particular identified expert.

In the conditions with the public rating system (***Rating*** and ***Exp+Rating***), consumers can choose to rate interactions with experts on a five-star rating scale after receiving their payoff in a given period.[18] This rating is shown to the respective expert at the end of the period.[19] Subsequently, ratings for each expert over all treated consumers are aggregated, averaged, and displayed to consumers. Consumers see these public ratings for all experts when they decide whether to interact and which expert to choose starting in period 5. In the condition without personal experience (***Rating***), as highlighted before, consumers cannot identify a particular expert and only see the public ratings. The public ratings of all experts are displayed to experts when they decide on the type of treatment and which price to charge in a given period (see Appendix C for the screenshots showing the feedback information provided to consumers and experts).

In addition to the main experimental conditions, we ran five further control conditions. Four control conditions under the healthcare frame were conducted to separate the role of expert competition, personal experience in the absence of expert competition, and private ratings.[20] These control conditions will be explained in more detail in the corresponding results sections whenever they are used to disentangle effects in the main conditions. A fifth control condition was implemented to check potential subject pool differences, since not all experimental sessions could be conducted in the same lab.[21]

3.3. Main outcome variables

The outcomes of interest are expert behavior, consumer decisions, and market efficiency. Table 2 below shows the payoffs and thus incentives for the stage game. Table 3 lists the key outcome variables and provides their description and measurement for the results section.[22]

[18] In essence we model a single-dimension rating system where consumers can give one overall rating for every interaction. Note that many physician platforms have adopted multidimensional rating systems where patients can rate multiple dimensions, like waiting times, office environment, or physician knowledge, which seems to enhance rating informativeness (Chen et al., 2018).

[19] We decided to inform the expert about the private rating to have full information provision about the rating to all participants irrespective of the history of play. To disentangle the effect of providing this information privately from the effect of the public disclosure of the average rating, see the results on the private rating condition in Appendix B.

[20] For a detailed description of the experimental conditions and the results see Appendix B.

[21] The detailed information on the experimental sessions is provided in the experimental protocol in Section 3.4.

[22] Appendix D lays out in more detail which expert and consumer behavior can be supported in equilibrium in the different experimental conditions.

Table 2: Payoffs in the stage game

	Expert provides and charges			
	q_L, p_L	q_L, p_H	q_H, p_L	q_H, p_H
Expert	$3-2=1$	$8-2=6$	$3-6=-3$	$8-6=2$
Consumer needs minor (q_L)	$10-3=7$	$10-8=2$	$10-3=7$	$10-8=2$
	Correct	**Overcharging**		(Overtreatment)
Consumer needs major (q_H)	$0-3=-3$	$0-8=-8$	$10-3=7$	$10-8=2$
	Undertreatment	**Undertreatment**		**Correct**

Expert behavior On the expert side, given the experimental setup and incentives, undertreatment and overcharging are the relevant expert decisions. Undertreatment is defined as the consumer needing the major treatment q_H, but the expert providing the minor treatment q_L. An expert might have incentives to do so since the costs for the major treatment are higher (6 ECUs versus 2 ECUs) and the expert can always charge the price of the major treatment (8 ECUs). In the results section, undertreatment will be reported in % of the expert-consumer interactions in which consumers need the major treatment. In terms of information, consumers can detect undertreatment in a period ex-post via their payoff, as the problem is not solved. In particular, if the expert charged p_H, the consumer payoff from undertreatment is -8 ECUs.

Overcharging is defined as the expert charging the price of the major treatment (p_H) while only providing the minor treatment to a consumer having a minor problem. In the results section, overcharging is reported accordingly in % of the expert-consumer interactions in which consumers need the minor treatment. In terms of information, consumers cannot infer ex post whether they have been overcharged, as they might have had a major problem requiring the major treatment charged at p_H. Thus, an expert can 'hide' behind a major treatment problem when overcharging.[23] If we had decided to impose verifiability of the provided treatment, the expert incentives would change from overcharging to overtreatment, as with verifiability of the treatment the expert can not just overcharge but has to overtreat. Please note that, in terms of consumer information, these are actually equivalent, so that the results for overcharging can be read as results for overtreatment under verifiability.

Consumer decisions On the consumer side, we record whether and with whom they choose to interact, and in the rating conditions whether they choose to provide a rating (captured by the variable feedback) and what the rating is (captured by the variable rating). Given the low outside option, except for very high risk aversion, consumers should always choose to interact,

[23] In principle, there is also scope for overtreatment, which would be providing the major treatment (q_H) and charging for it to a consumer with a minor problem, but overtreatment is strictly dominated by overcharging for the parametrization. In particular, instead of providing the major treatment with costs of 6 ECUs, for a consumer with a minor problem, the expert can always only provide the minor treatment (costs of 2 ECUs) and just (over)charge for the major treatment.

which is intentional in this study to mimic credence goods markets realistically. Our main focus of consumer decisions will therefore be the ratings themselves and the choice of experts over the periods.

Market outcomes We use two measures of market outcomes, overall market efficiency and consumer surplus. Market efficiency is driven by interaction (allowing surplus generation) and whether there is undertreatment, as undertreatment does not generate consumer value. Given our parametrization, we expect high levels of interactions, such that market efficiency is primarily determined by undertreatment. We normalize market efficiency, with 0% for no interaction and 100% for an interaction with the correct treatment. Consumer surplus incorporates the prices paid by consumers and is thereby influenced by overcharging, which is not the case for market efficiency. Consumer surplus is reported in absolute values.

3.4. Experimental protocol

The main experimental conditions shown in Table 1 were conducted using two market frames: a neutral and a healthcare frame. The sessions with the healthcare frame were conducted in the laboratory for experimental economic research at the University of Innsbruck. The sessions with the neutral frame were conducted in the laboratory for experimental economic research (experimenTUM) at the Technical University of Munich.[24] In each condition, we ran our experiment with 48 subjects, corresponding to 6 independent markets.[25] Regarding characteristics, the TUM sample has a slightly higher share of men (52% compared to 48%), and participants are on average 2.9 years older. To check for sample differences, the neutral frame treatment ***Rating-N*** was replicated in Innsbruck (see Section A.2 for the sample comparison). Overall, including the control conditions, 616 subjects participated.[26] All sessions were run computerized using z-Tree (Fischbacher, 2007) and participants were recruited using hroot (Bock et al., 2014). The project was approved by the internal review board of the University of Innsbruck. To ensure our target attendance of 24 participants (some sessions were run with 16 participants only), we invited 30 people to each session, however, dismissed all but 24 participants before starting the experiment. Those who did not get the chance to participate received a show-up fee of 4 Euros. At the beginning of each session, we explained the market setup to the participants, following a standardized protocol. An experimenter presented brief instructions to all

[24] To differentiate the two frames when necessary, the healthcare (neutral) frame conditions are labeled as ***Baseline-H*** (***Baseline-N***), ***Rating-H*** (***Rating-N***), ***Experience-H*** (***Experience-N***), ***Exp+Rating-H*** (***Exp+Rating-N***), respectively. When the extensions ***-H*** and ***-N*** are omitted, it implies the combined representation of results from both frames.

[25] The only exception is the neutral framing condition ***Exp+Rating-N***, which includes only 40 subjects due to recruitment challenges.

[26] 432 in Innsbruck, and 184 in Munich. There are more participants in Innsbruck as the additional control conditions were conducted in Innsbruck.

Table 3: Main outcome variables

Expert behavior	**Definition**	**Measurement**	**Notes**
Undertreatment (**UT**):	Consumer needs major treatment q_H, but expert provides minor treatment q_L.	As % of the expert-consumer interactions in which consumers need the major treatment.	Consumer can detect undertreatment ex-post in a given period by a low payoff (-8 ECUs if experts charged p_H).
Overcharging (**OC**):	Expert charges price of major treatment (p_H) but provides minor treatment q_L to a consumer needing only the minor treatment q_L.	As % of the expert-consumer interactions in which patients need minor treatment.	Cannot be identified ex-post by the consumer via payoff.
Overtreatment (**OT**):	Consumer needs minor treatment q_L, but expert provides major treatment q_H.	As % of the expert-consumer interactions in which consumers need minor treatment.	For the expert, overtreating is dominated by overcharging.
Consumer decisions			
Interaction:	At the beginning of every period, consumers decide whether to visit an expert.	Relative frequency of consumer-expert interactions.	Given the chosen low outside option, consumers should prefer to interact even when undertreated for the major problem unless they are strongly risk averse.
Choice of expert:	Which among the four experts in a market is chosen		Expert can be chosen by consumers according to ID in experience conditions and by rating in the rating conditions
Feedback:	After every interaction with an expert, consumers decide whether they want to give feedback in the rating conditions.	Relative frequency of giving feedback calculated as $\frac{\#\ of\ ratings}{\#\ of\ interactions}$.	
Rating:	Given that consumers give feedback, the rating given to expert on a scale from zero to five stars.	Average (expert) rating calculated as $\frac{sum\ of\ ratings}{\#\ of\ ratings}$.	In a given period, the rating is calculated using all ratings up to this period. We distinguish public and private ratings: The public rating of an expert uses ratings from all consumers, a private rating of a consumer for an expert is using only the ratings of this consumer.
Market outcomes			
Market efficiency (**EFF**):	Overall realized market surplus	per possible interaction: 0% if there was no interaction, 100% if the patient was treated correctly, 0.25 (0.67) if the patient was undertreated (overtreated). Aggregated by averaging.	
Consumer surplus (**CS**):	Overall realized consumer surplus	in absolute value, per possible interaction: consumer value of provided treatmnent - charged price , or outside option. Aggregated by averaging.	

subjects, covering the main features of the decision problem. Afterward, we asked subjects to read detailed instructions of the game and to answer a set of incentivized control questions (see Appendix E for the instructions and control questions). Once all subjects correctly answered the control questions, they were informed of their randomly assigned roles and played the credence goods game for 16 periods. At the end of the game, subjects participated in an individual risk preference task, a dictator game, a lying task, and a trust game. Finally, participants filled out a questionnaire (see Appendix F for the additional instructions and the questionnaire). The payment subjects received at the end of the session consisted of their profits from the credence goods game (4 randomly selected periods), one randomly selected additional task, and a lump sum payment for answering the questionnaire. Subjects earned 27.10 euros on average and sessions lasted approximately 105 minutes. For an overview of the sample characteristics Table A1 and Table A2 in Appendix A provide descriptive statistics on the background information collected by experimental conditions for the healthcare frame, respectively, the neutral frame.

3.5. Predictions and research questions

In this section, we discuss the main theoretical predictions and formulate our research questions. The analysis is based on the assumptions of rationality and, for simplicity, risk neutrality of experts and consumers. The benchmark is condition ***Baseline*** in which consumers can choose an expert, but cannot use information about past expert behavior in their expert choice as they can neither identify a given expert nor use rating information. This condition implements repeated first-time expert-consumer interactions where consumers choose between experts about whom no information is available.

The outside option of remaining untreated in ***Baseline*** and all other conditions is such that a consumer prefers to interact: Even when undertreated in the case of the major problem, and always charged the high price p_H, the expected payoff from interacting (-3) is higher than the outside option (-4). This is different from other credence goods experiments and reflects the important fact that in many credence goods markets, consumers are better off seeing the expert in expectation. Without other-regarding preferences of experts, the unique equilibrium in the stage game is that consumers always interact and experts always undertreat and overcharge. As reputation-building of experts is not possible, the equilibrium over all periods is the repeated stage game equilibrium. If experts have social preferences such as altruism and efficiency concerns, they might however not always undertreat/overcharge. The results from ***Baseline*** allow to have an aggregate measure of these social preferences of experts given the market set-up.

Conditions ***Rating***, ***Experience*** and ***Exp+Rating*** then allow consumers to use information about past expert behavior, albeit through different channels. The basic consumer information

takes the following form: consumers observe their payoff from an interaction with an expert, and can infer whether this expert undertreated them. Furthermore, if they are only charged price p_L (and not undertreated), they can even infer that the expert did not overcharge them. We will henceforth call either of these two basic forms of information (no undertreatment, no overcharging) a positive consumer experience.

In ***Experience***, consumers can identify experts and have their own past experiences as information about the behavior of an identified expert. Punishing (rewarding) an expert by not visiting (re-visiting) the expert based on this personal experience information then allows for reputation equilibria to exist in condition ***Experience***. In these, experts build up a reputation for quality in early periods based on the following consumer strategies: consumers stay with an expert for which their belief about a positive experience is sufficiently high in early periods. Conversely, negative personal information leads to punishment by not (re)choosing the corresponding expert. In late periods, experts who provided a positive experience in early periods milk their reputation and are rewarded by consumers staying with them and thereby allowing them to make (high) profits at/until the end. The reputation incentives for experts are thus a back-loaded remuneration, and this works as consumers have earlier period personal experience information.

In ***Rating***, reputation for quality equilibria may exist as well, albeit through a different channel. Consumers cannot choose experts based on their own experience, but they have access to aggregated, indirect information from other consumers' experiences. Interpreting this information requires a belief about the rating strategies of other consumers. If they are such that consumers believe that other consumers give a high rating (on the 0-5 star scale) when they had a positive consumer experience, then a higher rating leads to a higher belief about the expert providing a positive consumer experience in early periods. The following consumer strategies may then sustain a reputation for quality equilibria: consumers give a high rating to an (unidentified) expert when they had a positive consumer experience, and in the next period choose an (unidentified) expert with a high rating. In late periods, for rewarding experts with high ratings in early periods, consumers continue to choose experts with the highest ratings in the reputation-milking phase, in order for these experts to keep their customers.

In terms of outcomes, the following types of a reputation for quality equilibria can be supported:[27] **Equilibria without undertreatment** in early periods, as well as **equilibria without undertreatment and without full overcharging** in early periods. These differ in whether consumers punish experts by not re-visiting them (***Experience***) or giving low ratings (***Rating***) in early periods only when they receive a negative payoff (undertreatment) or also when they

[27] The structure of these equilibria is described below. Appendix D shows the construction. There is a multiplicity of equilibria with e.g. different switching periods between reputation-building and reputation-milking. Of course, no reputation equilibria with full undertreatment and overcharging exist as well.

are charged the high price (p_H). The latter case is more complex as consumers cannot distinguish between being overcharged or not: p_H is not overcharging when the consumer has a major problem. Nevertheless, punishing when charged p_H can sustain equilibria without full overcharging in which experts undercharge in early periods.[28]

Thus, although via different channels, reputation equilibria may exist in both ***Experience*** and ***Rating*** as well as the combination (***Exp+Rating***).[29] Whether they emerge and are more likely to prevail in ***Experience*** or ***Rating***, or require the combination of both, is an empirical question and the motivation for taking the problem to the lab. The indirect information about expert behavior in credence goods markets from ***Rating*** might thereby be perceived as noisier and less reliable, as it depends on the rating strategies of other consumers. In particular, the belief about expert behavior (reputation) depends on the beliefs about other consumers' rating behavior. Conversely, the public rating might also be considered as containing more information and being more salient compared to personal experience. The 2×2 experimental design can provide results both on the effectiveness of a public rating system and whether public rating and personal experience are complements or substitutes for reputation-building. The focus of the analysis will thus be on the following research questions:

Research Question 1 *What is the impact of public rating information on expert behavior and market efficiency in markets without personal experience? (**Baseline** vs. **Rating**)*

Research Question 2 *Is public rating information a good substitute for personal experience information in terms of market outcomes? (**Rating** vs. **Experience**)*

Research Question 3 *Is public rating information a complement to personal experience information in terms of market outcomes? (**Experience** vs. **Exp+Rating**)*

Research Question 4 *Do consumers react less strongly to public ratings than to personal experience information? (Analysis of consumer decisions in **Exp+Rating**)*

[28] Compared to equilibria without undertreatment, those with additionally less than full overcharging have a shorter period of good expert behavior and a longer period of rewards as expert profits when undercharging are substantially lower.

[29] As the experimental design is one of pure moral hazard, the discussion of reputation equilibria above was focused on reputation for quality equilibria, and not on reputation for type equilibria. Of course, additionally, the logic of reputation for (nonfraudulent) type might be present, with information on past expert behavior helping to select good types. It is not the focus of the present experiment to disentangle the different reputation equilibria. The mechanisms in terms of the role of consumer information and decisions for reputation-building are, in any case, analogous: Information about past expert behavior, either via own experience or public ratings, can channel the consumer choice of expert.

4. Results

We begin by examining the aggregate results in Section 4.1 to answer the main research questions. To better understand the dynamics behind the aggregate results, and to confirm whether and how these results pertain to reputation-building, we will analyze ratings and consumer decisions in more detail in Section 4.2.

We present and discuss the results for the main conditions, combining data from both the healthcare and neutral frames. The decision to pool the data is based on the consistency of key findings across both frames. Differences between frames will be discussed in parallel, detailed results for each frame individually are available in Appendix A.[30]

Table 4 reports the aggregate results averaged over markets and periods, with the corresponding non-parametric tests for the relevant experimental condition comparison. To complement the non-parametric results, Table 5 reports on the results from multilevel mixed-effects probit and linear regressions.[31] We ran two different models: The first model shows the effect of our experimental conditions when controlling only for time trends. In the second model, we control for economic preferences and personal characteristics relevant in a credence goods setting by adding experimental measures for social preferences, lying, trustworthiness as well as measures for personality traits alongside the standard socio-demographic covariates. Figure 1 displays the main results averaged over markets throughout the 16 periods.

4.1. The effects of a public rating system

Column 1 of Table 4 shows that there is substantial undertreatment and overcharging in a market without either personal experience or public rating (***Baseline***): experts undertreat their consumers in 52.4% of all cases and overcharge them in 85.8% of all cases. Market efficiency, which is determined by undertreatment and interactions, is at only 76.8%.[32] As a benchmark, full undertreatment with full interaction would lead to a market efficiency of 62.5%, as there is no efficiency loss for consumers with the minor problem, and only no interaction could lead to a market efficiency below 62.5%.

The introduction of a public rating system (Column 2 of Table 4) leads to a sharp and significant decline in undertreatment, dropping from 52.4% in ***Baseline*** to only 7.2% in ***Rating***.

[30] For a detailed comparison of the results in condition ***Rating-N*** between participants from the University of Innsbruck and the Technical University of Munich, see Section A.2.

[31] Multilevel mixed-effects models are designed specifically to account for dependencies between observations on different hierarchical levels. In our case, we use a three-level mixed-effects model to account for the dependency of observations at the subject and/or market levels.

[32] In the baseline condition with a healthcare (neutral) frame, ***Baseline-H*** (***Baseline-N***), experts undertreat their consumers in 64.7% (40.1%) of all cases and overcharge them in 92% (79.6%) of all cases. Market efficiency is at 70.7% (82.9%) (see Table A3 and Table A4 in Appendix A).

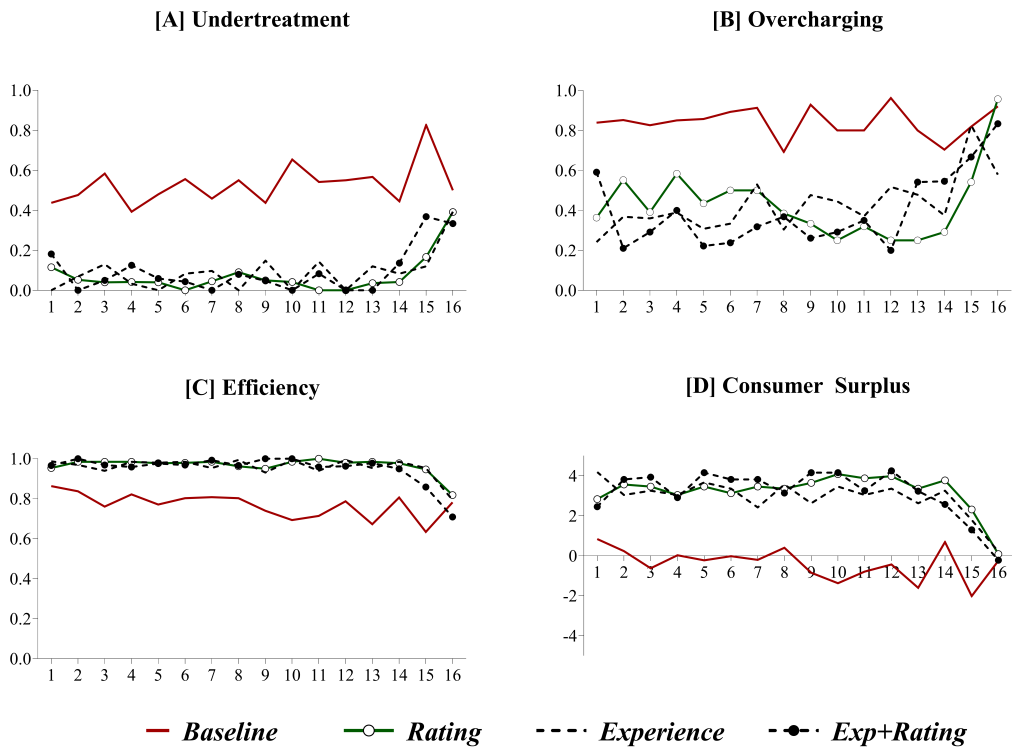


Figure 1: Rate of undertreatment, overcharging, efficiency, and consumer surplus by experimental conditions.

This induces a significant increase in market efficiency from 76.8% to 96.6%. Furthermore, overcharging also significantly decreases from 85.8% to 43.9%.[33] This is a particularly interesting finding, as overcharging—contrary to undertreatment—cannot be directly observed by the consumer. Despite the possibility of hiding behind the probability of a major problem for the treatment of which the high price can be reasonably charged, the disciplining effect of ratings reduces this overcharging behavior. The reduction in both undertreatment and overcharging leads to a substantial increase in consumer surplus.

These as well as all the following results from the nonparametric analysis on experimental condition comparison are confirmed in the regression analysis. The results on the effect of a public rating system are also validated in the healthcare and neutral frame conditions separately, with the effects being quantitatively less strong in the neutral frame. The latter is a result of a lower undertreatment rate in the neutral frame ***Baseline-N***, at 40.1%, compared to 64.7% in the healthcare frame ***Baseline-H***, somewhat contrary to intuition about the effect of a health care frame (see Table A3 and Table A4 in Appendix A).

Result 1 *Introducing a public rating system into a credence goods market (without personal experience) significantly decreases both undertreatment and overcharging and significantly increases*

[33] These results also hold when restricting the comparison between conditions to the first eight, respectively the last eight periods.

Table 4: Overview of results (means)

	Markets without personal experience		Markets with personal experience		*p-values of MWU*[1]		
	Baseline	***Rating***	***Experience***	***Exp+Rating***	***Baseline vs Rating***	***Rating vs Experience***	***Experience vs Exp+Rating***
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Expert behavior							
Undertreatment (in %)	52.42	7.21	9.22	9.08	0.000	0.619	0.774
Overcharging (in %)	85.83	43.90	42.62	39.67	0.000	0.744	0.684
Overtreatment (in %)	1.05	2.62	4.06	3.15	0.576	0.511	1.000
Consumer decisions							
Interaction (in %)	95.96	99.48	99.87	98.44	0.362	0.466	0.011
Feedback (in %)	-	94.99	-	90.80			
Star-rating	-	3.77	-	3.65			
Market outcomes							
Efficiency (in %)	76.80	96.55	95.81	95.05	0.000	0.744	0.639
Consumer Surplus (in ECUs)	-0.39	3.21	2.93	3.17	0.000	0.340	0.477
Observations	96	96	96	88			

Note: We analyze twelve independent markets in every experimental condition except for ***Exp+Rating***, where we only have eleven markets. In each market, four consumers and four experts interact. The experimental conditions are: ***Baseline***, ***Rating***, ***Experience***, and ***Exp+Rating***. Please refer to Section 3.2 for a description of the main experimental conditions. See Table 3 for a description of the outcome variables.
[1] Mann-Whitney U-tests for pairwise differences between conditions with matching groups of 8 subjects as one independent observation. p-values are adjusted for the small sample size, using Fisher's exact test.

consumer surplus and market efficiency.

One important question is whether it is the reputation-building mechanism via ratings that drives this result. We will analyze ratings and rating dynamics in more detail in the next section.

In the condition with personal experience but without a public rating system, undertreatment is at only 9.2% and overcharging is at 42.6% (Column 3 of Table 4). Compared to ***Baseline***, we find a significant decrease in both undertreatment and overcharging which leads to a significant increase in consumer surplus and efficiency (p-value MWU: <0.01 all). When comparing ***Rating*** to ***Experience***, we find that aggregate results on expert behavior are very similar: The undertreatment rate at 7.2% and the overcharging rate at 43.9% in ***Rating*** is almost the same as that in ***Experience***. There is no significant difference in either consumer surplus or market ef-

Table 5: Average treatment effects

	Undertreatment		**Overcharging**		**Efficiency**	**Consumer Surplus**
	(1)	(2)	(3)	(4)	(5)	(6)
Levels in ***Baseline***	0.538	0.528	0.878	0.864	0.768	-0.391
	(0.071)	(0.075)	(0.045)	(0.037)	(0.044)	(0.544)
	Marginal treatment effects					
Rating	-0.410***	-0.405***	-0.366***	-0.389***	0.198***	3.603***
	(0.074)	(0.080)	(0.060)	(0.053)	(0.045)	(0.587)
Experience	-0.388***	-0.388***	-0.381***	-0.406***	0.190***	3.320***
	(0.077)	(0.082)	(0.061)	(0.059)	(0.045)	(0.575)
Exp+Rating	-0.385***	-0.366***	-0.448***	-0.462***	0.182***	3.561***
	(0.077)	(0.085)	(0.061)	(0.055)	(0.046)	(0.577)
Period	+***	+***	+***	+***	-***	-***
	Additional Games					
Amount donated to charity		not sig.		-*		
Liar (yes)		not sig.		not sig.		
Trustworthiness		not sig.		not sig.		
Covariates		✓		✓		
	p-values from post-estimation Wald-Test					
Rating vs ***Experience***	*0.467*	*0.632*	*0.793*	*0.745*	*0.516*	*0.330*
Rating vs ***Exp+Rating***	*0.401*	*0.293*	*0.146*	*0.212*	*0.270*	*0.887*
Experience vs ***Exp+Rating***	*0.926*	*0.636*	*0.231*	*0.374*	*0.610*	*0.372*
Observations	1475		1486		3008	3008
Number of Groups	47		47		47	47

Note: The table presents results from multilevel models with random effects at the market and individual levels (undertreatment & overcharging: columns 1-4) or at the market level for market efficiency (column 5) and consumer surplus (column 6). See Table 3 for a description of the outcome variables. We report effects as marginal effects, calculated as differences in the expected probabilities between the experimental condition in question and the baseline condition. All regressions include time trends. **Covariates**: Gender, age, BIG 5 personality traits (extraversion, agreeableness, conscientiousness, neuroticism, openness) measured with a 10-item BIG 5 questionnaire, whether the participant is a business/economics major, self-reported frequency of practicing religion, number of physician visits in the past 12 months, an indicator for experience with incorrect physician behavior, an indicator for experience with physician recommendations, relative school performance as a proxy for IQ, a measure for altruism (the amount donated to charity in a dictator game), an indicator whether the participant is classified as a liar (if reporting 4 or more correct dice rolls out of 12 in a lying task), and trustworthiness measured in a standard trust game. Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ficiency.[34] Thus, overall, a public rating system appears to implement similar market outcomes as a market in which consumers can rely on personal experience information about experts.

[34] This is confirmed in both frames separately (see Table A3, Table A5, Table A4, and Table A6 in Appendix A). In the healthcare frame, the difference in overcharging between ***Rating-H*** and ***Experience-H*** is slightly higher (MWU: $p = 0.093$), however, it is not statistically significant in the regression analyses when additional covariates are controlled for (see Table A5).

Result 2 *There is no significant difference in market outcomes between **Rating** and **Experience**. With respect to overall market outcomes, the public rating system is a good substitute for personal experience information in the studied credence goods markets.*

We now turn to the question of the effect of a public rating system when consumers can rely on information about expert behavior from their personal experience with the expert, in particular, whether personal experience and public rating information is complementary. Columns 3 and 4 of Table 4 show that markets with personal experience, both without and with a public rating system, have a low level of undertreatment (9.2% and 9.1% respectively) and moderate levels of overcharging (42.6% and 39.7% respectively). For all variables of expert behavior as well as consumer surplus and efficiency, there are no statistically significant differences between ***Exp+Rating*** and ***Experience***. Similarly, using Result 2, the vice versa case for adding personal experience information to a public rating system is the same. One observation regarding these results, however, is that the potential to improve market outcomes by supplementing personal experience with public rating information (or vice versa) is limited, as undertreatment is already close to the First Best level in ***Experience*** and ***Rating***. Taken together, we do not find a complementarity of public rating and personal experience information in our experiment, but this can be explained by an already high level of market performance in a market with either personal experience or public rating, which reduces the scope for complementarity.[35]

Result 3 *Introducing a public rating system into a market in which consumers have personal experience with experts neither improves (nor worsens) market outcomes. We find no complementarity between public rating and personal experience information with respect to overall market outcomes.*

In addition to the differences between conditions, the coefficient for the time trend in Table 5 shows that undertreatment and overcharging increase, while market efficiency decreases over time. Moreover, the regression results in Columns 2 and 4 in Table 5 together with the regression results for both frames separately (Appendix A) show no consistent patterns for the economic preference measures from the additional games.[36]

[35] These results hold across both frames with one exception: In the neutral frame we find a significant difference between ***Rating-N*** and ***Exp+Rating-N*** for undertreatment and a weakly significant difference for the efficiency (see Table A6).

[36] In the health frame, there is an association between undertreatment, overcharging and the amount donated to charity: participants who are willing to give more money to a charity in a dictator game engage less often in both, undertreatment (at the 1% level) and overcharging (at the 10% level). In the neutral frame, trustworthiness shows a negative association with undertreatment at the 10% level (see Table A5 and Table A6 in Appendix A).

4.2. Consumer ratings and expert selection

In this section, we explore consumer behavior with a special emphasis on the ratings and expert choice. In ***Rating***, the large majority (95%) of interactions are rated and the average rating is 3.8 stars. Similarly, in ***Exp+Rating*** 90.8% of interactions are rated with an average rating of 3.7 stars (see Table 4). These average star ratings hide a substantial differentiation by the consumer payoff. Figure 2 shows the average star rating by consumer payoff for the two experimental conditions with ratings.

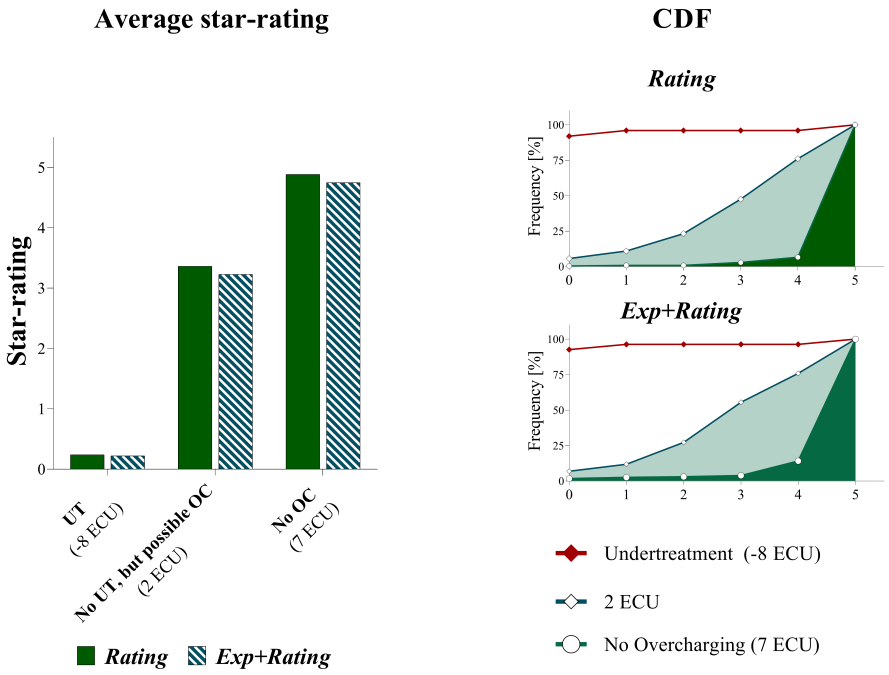


Figure 2: Rating behavior of consumers. On the left side, we see the mean ratings for each of the possible payoffs of consumers. The right side shows the cumulative distribution function (CDF) of given star ratings, separately for possible payoffs of consumers. If consumers are undertreated, the payoff is -8 ECUs, whereas if they have a minor problem and are treated appropriately, the payoff is 7 ECUs. In the case of a minor problem and appropriate treatment but overcharging, or in the case of a major problem and appropriate treatment with charges, the payoff is 2 ECUs.

When experiencing a negative payoff (-8 ECUs) in a period, consumers can infer that they were undertreated in this period. This leads to a rating of 0 stars, in fact, this was the case for almost all interactions except for five, in which consumers gave a higher rating. Thus, undertreatment, which can be observed ex-post, leads to an unambiguous punishment that is symmetric across consumers (see also Table 6).

When receiving a payoff of 7 ECUs, consumers can infer that they were neither undertreated nor overcharged. In that case, 89.8% of interactions were rated with five stars. The most interesting part is the rating given for a payoff of 2 ECUs: This payoff is generated when the

consumer either had a major problem and was appropriately treated and charged or when the consumer had a minor problem and was overcharged. Thus, the expert can 'hide' behind a major problem and overcharge in case of a minor problem. The distribution of ratings for this case is more dispersed, with ratings in the two conditions ranging from 0 (6.2%) to 5 (24%) and a median rating of 3 stars.[37]

Table 6: Ratings response

	Star rating
Predicted star-rating if patient payoff is 7 ECUs	4.76
	Marginal effects if ...
... payoff is 2 ECUs	-1.420***
	(0.111)
... payoff is -8 ECUs	-4.363***
	(0.229)
Observations	1355
Number of groups	23

Note: For this analysis, only the treatments ***Rating*** and ***Exp+Rating*** are considered. The table presents the marginal treatment effects of multilevel models with random effects at the market and individual levels. The dependent variables are star ratings following an interaction with an expert. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Apart from ratings, it is essential for reputation as an indicator of quality to be effective that consumers visit experts whom they anticipate will provide a high quality of care (positive consumer experience). As highlighted in Section 3.5, the channel is staying with an expert in the personal experience conditions, and going to experts with the highest ratings in the rating conditions, conditional on symmetric strategies by the other consumers. The latter— symmetric strategies where undertreatment is clearly punished with a bad rating and no undertreatment/no overcharging is clearly rewarded with a good rating— seems to be the case. In ***Baseline***, on the contrary, consumers do not have information that can (re)direct them to experts for which they can expect high quality.

Figure 3 shows the frequency of a change in expert by the consumer payoff. The line corresponds to the expected frequency associated with random assignment among the 4 experts in a market (75%). Figure 3 nicely illustrates that in ***Baseline***, as experts cannot be identified, consumers cannot change intentionally and thus cannot provide incentives via their expert visit decisions. In ***Rating***, while consumers cannot identify experts, they can provide high ratings when they receive a high payoff and decide to choose best-rated experts which often implies staying with the same expert, and this can explain the lower frequencies of change in ***Rating***

[37] The results for the healthcare and neutral frame conditions show similar results in Figure A3 and Figure A4, respectively Table A7 and Table A8 for the regression results.

compared to ***Baseline*** for consumer payoffs 2 ECUs and 7 ECUs. However, and in line with intuition, the reaction of consumers is strongest in markets with personal experience where experts are identified. Note that this latter result is mainly driven by the healthcare frame conditions as shown in Figure A5. In neutral frame conditions, consumers' reactions to a positive consumer experience is similar in both markets with personal experience and those without (see Figure A6).

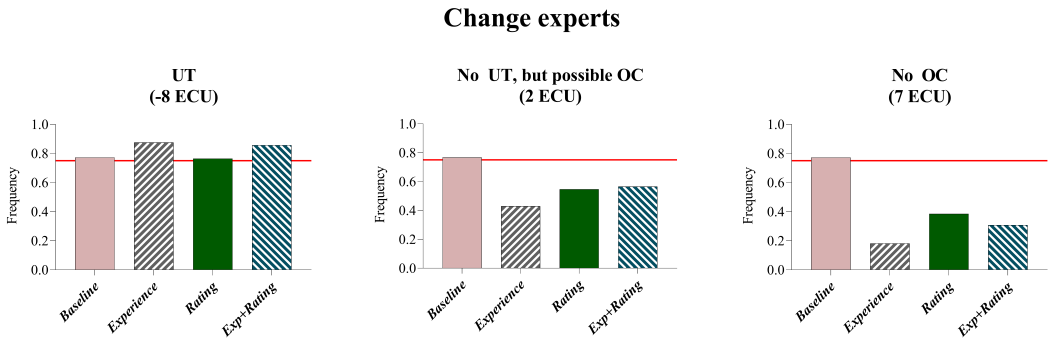


Figure 3: Frequency of a change in expert by realized consumer payoff in a given period.

Table 7: Associations between payoffs of consumers and their decision to change the expert

	Change expert
Frequency of change if consumer-profit is 7 ECUs	0.25
	Marginal effects if ...
... patient-profit is 2 ECUs	0.238***
	(0.032)
... patient-profit is -8 ECUs	0.548***
	(0.056)
Observations	1371
Number of groups	23

Note: For this analysis, only the treatments ***Experience*** and ***Exp+Rating*** are considered. The table presents results from a three-level model with random effects at the market and individual levels. The dependent variable is a binary indicator of whether a consumer changed the expert. We report effects as marginal effects, calculated as differences in the expected probabilities between the payoff in question and the maximum profit of 7 ECUs. Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The regression analysis shown in Table 7 confirms that the decision to change the expert depends strongly on the consumer payoff, with behavior consistent with predictions for the reputation equilibria. While we show the frequency of change also for ***Baseline*** and ***Rating*** in Figure 3, the results on the decision to change reported in Table 7 are based only on conditions ***Experience*** and ***Exp+Rating*** to account for the fact that consumers can only fully intention-

ally leave a given expert in these two experimental conditions.[38]

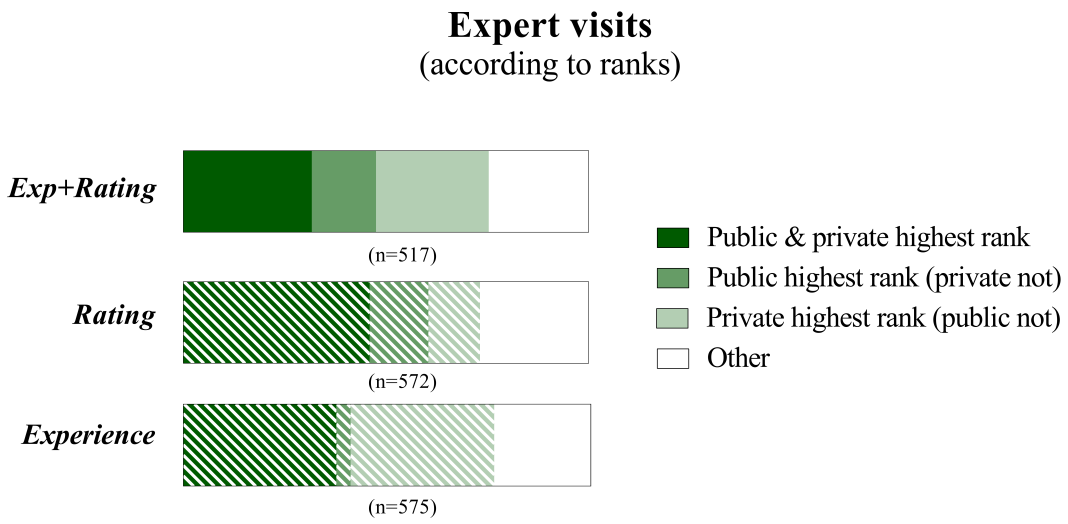


Figure 4: Distribution of consumers' realized expert visits according to relative private and public rankings of experts. Results for ***Rating***, and ***Experience*** are crosshatched as consumers do not have full information on both private and public rankings in these conditions.

Figure 4 shows expert visits depending on their ranking with respect to both public and private ratings, and Table 8 provides the corresponding regression results. The private rating of a consumer is the average rating that the consumer gave to the expert up to the corresponding period. We distinguish four categories of visits: whether the expert visited was the highest ranked in both public and private rating, had the highest public but not private rank, the highest private but not public rank, or did not fall in either of the previous categories (other). For interpretation, it is important to note that both public and private ratings are only explicitly available to a patient in ***Exp+Rating***. For this reason, we speak of realized expert visits but not choice/selection. In ***Rating***, the private rating is implicit as consumers cannot attribute it to a given expert. For the condition ***Experience*** in which consumers do not rate experts, we have constructed a hypothetical private and public rating of experts given their choices based on the corresponding average ratings for the same choices in ***Rating***.[39]

The first observation from Figure 4 is that the majority of consumers selects the publically best-rated experts in ***Rating*** and ***Exp+Rating***. Interestingly, in all conditions, visits with experts

[38] The regression results on the effect of the consumer payoff on the decision to change the expert for the healthcare frame are presented in Table A9 and for the neutral frame in Table A10 in Appendix A.

[39] In Figure A7 the same results are shown for the healthcare frame. Note that here we also display the result for the control condition ***Exp+Rating-Priv-H***, which is the same as ***Experience-H*** except that consumers give a private rating to experts which can be used as a private rating and aggregated to a public rating. It is reassuring to see that the distribution of consumers' expert visits according to ranks looks almost identical in both conditions, in ***Experience-H*** in which hypothetical ranks are reconstructed and ***Exp+Rating-Priv-H*** in which ranks are based on actual ratings.

Table 8: Expert visits according to rank

	Public and private best	***Public best (private not)***	***Private best (public not)***	***Other***
	(1)	(2)	(3)	(4)
Levels in ***Rating***	0.460	0.145	0.128	0.267
	Marginal treatment effects			
Experience	-0.084***	-0.110***	0.225***	-0.031
	(0.029)	(0.017)	(0.024)	(0.026)
Exp+Rating	-0.143***	-0.014	0.151***	-0.022
	(0.029)	(0.022)	(0.024)	(0.026)
	p-values from post-estimation Wald-Tests			
Exp+Rating vs ***Experience***	0.042	0.000	0.008	0.725

Note: The table presents results from a multinomial logistic regression. We report the predicted frequencies of choosing experts based on public and private ranks in ***Rating*** and the difference between ***Rating***, ***Experience***, and ***Exp+Rating*** as marginal effects. E.g. consumers in ***Rating*** choose the private (but not public) best-rated expert in 12.8% of cases (column 3), while consumers in ***Experience*** do so significantly more often, in 35.3% of the cases.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

that had both the top public and private rating ranks are the most frequent.[40] Thus, even though both private and public ratings are not available in all conditions, the feedback information available in the respective condition effectively channels consumers to the individually and publicly best-rated experts. Furthermore, Table 8 shows that the shares of expert visits for which the private rank but not public rating rank is highest are significantly higher in all conditions with personal experience compared to ***Rating***. Similarly, the shares of expert visits with the best public but not private rank is lower in ***Experience*** compared to ***Rating***.

An important question is which type of information is more relevant for consumers' selection of experts when they have both personal experience information and public rating information available. To interpret the relative importance of personal experience vs. public rating information in expert *choice*, we look at ***Exp+Rating*** in more detail. Figure 5 shows the distribution of selected experts by the spread in public-private rank (left) and rating (right).[41] Both distributions are left-skewed (left: -0.409 , p < 0.01[42]; right: -0.795 , p < 0.01), revealing that

[40] For the frames separately, the share of visits with experts that had both the top public and private ratings is slightly higher in ***Rating-N*** than in ***Rating-H*** but lower in ***Experience-N*** than in ***Experience-H*** (see Figure A7 and Figure A8 in Appendix A).

[41] The analogous results for the healthcare and neutral frame are shown in Figure A9 and Figure A10 in Appendix A.

[42] For this analysis, we exclude the 209 visits where ranks were equal and only considered interactions where there was a discrepancy between the private and the public ranking.

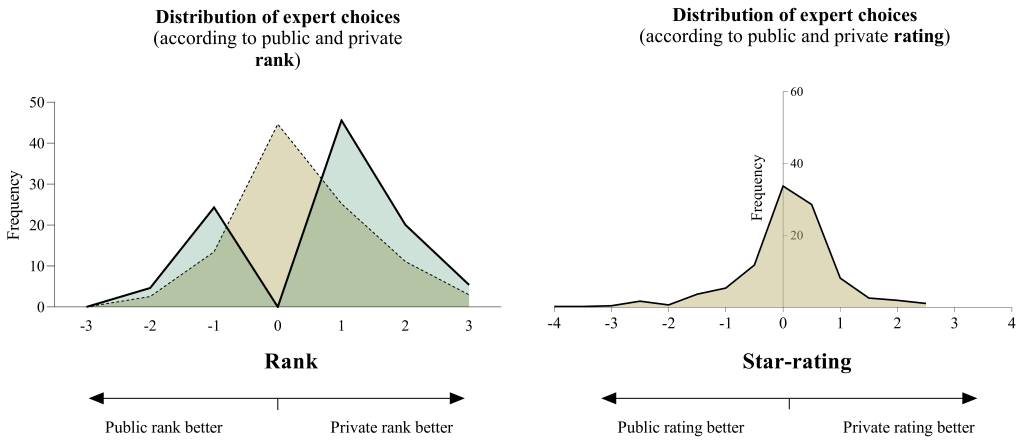


Figure 5: Distribution of selected experts by the spread in public-private rank (left) and rating (right). For this figure, only the treatment ***Exp+Rating*** is considered. **Left**: The dashed line shows the distribution of choices according to ranks including equal ranks. We observed 468 interactions where consumers chose an expert for whom they had both, private experiences, and a public rating. Of those, consumers selected an expert with equal ranks in 44.6% (209 interactions). The solid line (shaded area) only shows the distribution of selected experts when there was a discrepancy between the private and the public ranking. Testing for normality reveals that the distribution is significantly skewed to the left (-0.409, $p < 0.01$). **Right**: we show the distribution of selected experts according to differences between the private and public average ratings (private average rating - public average rating). Hence, positive (negative) numbers indicate that the expert had a better private (public) rating. The distribution is significantly skewed to the left (-0.795, $p < 0.01$).

consumers, when selecting experts put more weight on their private experience than on the public rating.

Result 4 *The majority of consumers select the best-rated experts in* ***Exp+Rating*** *and* ***Rating****. When there is a discrepancy between the private experience and the public ratings, consumers seem to put more weight on information from private experiences when choosing experts in* ***Exp+Rating****.*

5. Discussion and conclusion

Online rating platforms have become increasingly common in recent years. Nevertheless, there remains a scarcity of studies that investigate the causal effect of public rating systems on market outcomes. In this paper, we experimentally investigate the effect of the prominent five-star rating system on expert behavior in a credence goods market. The experiment thereby uses two frames: A neutral frame and the frame of one of the most important credence goods markets, health care. Importantly, the key results are fully supported in both framings suggesting that the results extend to other credence goods markets.

In the experiment, we distinguish between two different market environments: Those in which consumers can base their expert choice on their own previous experience with a particular expert, and those in which this is not the case. The results show that though consumers cannot judge all relevant quality aspects even ex-post, a public rating system significantly improves market efficiency and consumer surplus. Even overcharging, which cannot be detected, decreases significantly. When it comes to expert behavior, ratings can influence it essentially in two ways—directly, through a signal sent by the consumer to an expert, and indirectly, through the reputational effect of those ratings. Our results suggest that the reputational effect is the driving force.

Furthermore, we find that a public rating system is a good substitute for personal experience information to enhance market outcomes. Market efficiency and consumer surplus are on the same levels in markets with a public rating system compared to personal experience markets. When consumers have both personal experiences as well as public ratings to base their decisions on, they tend to prioritize the former over public ratings. This aligns with previous research in experience goods suggesting that public rating systems are particularly helpful for those interacting for the first time, like tourists and travelers (Fang, 2022), but do not offer additional benefits when market participants can rely on personal relationships (Cai et al., 2014).

Considering ratings, we see that consumers use them effectively to reward or punish experts, which allows reputation equilibria to emerge. We find that consumers symmetrically punish experts with a *zero* rating when being undertreated, give low ratings (albeit more dispersed) when being charged the high price, and reward—again symmetrically—experts for which they know that they did neither undertreat nor overcharge with a five-star rating. These rating strategies appear to be well understood by all market participants as they then strongly direct subsequent expert choice.

In our experiment, consumers observe ex post whether they have been undertreated. This might change with the same expert in the next period, so the fact that undertreatment in a given period is observable ex post does not make the service of the expert an experience good, but this dimension has somewhat the flavor of an experience characteristic. One extension for the experimental analysis would be to introduce diagnostic errors, so that even undertreatment provides less precise information about expert behavior, as it might be due to an error. While we did not implement this extension in the current experiment, our results regarding overcharging can give an indication for this case: Undertreatment, when it might be due to a diagnostic error, is quite similar to overcharging in our setting in terms of the inferences a consumer can make from observations of expert behavior.

Furthermore, by not imposing verifiability of the treatment, the relevant incentive for experts in our experiment pertains to overcharging and not overtreatment. This might be questioned as

notably for healthcare markets, overtreatment is often considered the more prominent problem compared to overcharging. Importantly, our results for overcharging can be interpreted as results on overtreatment: Had we imposed verifiability, the incentive would have switched from overcharging to overtreatment, with all other relevant mechanisms in terms of consumer information and inference and consequently reputation-building remaining unchanged.

Despite the potential benefits of public rating mechanisms, there are also concerns about their accuracy and reliability. Ratings are affected by various factors other than the genuine quality of the service provided (Doing-Harris et al., 2016; López et al., 2012; Okike et al., 2016), in particular in credence goods markets. Compared to the experimental setup, ratings tend to be more subjective and the information provided about the quality of expert decisions becomes less reliable. One concern is that experts might compensate low quality in some dimensions by higher quality in other dimensions that can be more easily judged in order to receive a better rating. Review fraud is another potential concern. Studying the resilience of public rating platforms in the face of growing levels of noise and reduced reliability of ratings is a vital area for future research.

References

Angerer, S., Glätzle-Rützler, D., & Waibel, C. (2021). Monitoring institutions in healthcare markets: Experimental evidence. *Health Economics*, 30(5). https://doi.org/https://doi.org/10.1002/hec.4232

Angerer, S., Glätzle-Rützler, D., & Waibel, C. (2023). Framing and subject pool effects in healthcare credence goods. *Journal of Behavioral and Experimental Economics*, 103, 101973. https://doi.org/https://doi.org/10.1016/j.socec.2022.101973

Anthun, K. S., Bjørngaard, J. H., & Magnussen, J. (2017). Economic incentives and diagnostic coding in a public health care system. *International Journal of Health Economics and Management*, 17(1), 83–101. https://doi.org/10.1007/s10754-016-9201-9

Ba, S. & Paul, A. P. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly*, 26(3), 243–268. https://doi.org/10.2307/4132332

Baker, L. C. (2010). Acquisition of mri equipment by doctors drives up imaging use and spending. *Health Affairs*, 29(12), 2252–2259. https://doi.org/10.1377/hlthaff.2009.1099. PMID: 21134927

Balafoutas, L., Beck, A., Kerschbamer, R., & Sutter, M. (2013). What drives taxi drivers? a field experiment on fraud in a market for credence goods. *The Review of Economic Studies*, 80(3), 876–891. https://doi.org/10.1093/restud/rds049

Balafoutas, L. & Kerschbamer, R. (2020). Credence goods in the literature: What the past fifteen years have taught us about fraud, incentives, and the role of institutions. *Journal of Behavioral and Experimental Finance*, 26, 100285. https://doi.org/10.1016/j.jbef.2020.100285

Balafoutas, L., Kerschbamer, R., & Sutter, M. (2017). Second-degree moral hazard in a real-world credence goods market. *The Economic Journal*, 127(599), 1–18. https://doi.org/10.1111/ecoj.12260

Bar-Isaac, H. & Tadelis, S. (2008). Seller reputation. *Foundations and Trends® in Microeconomics*, 4(4), 273–351. https://doi.org/10.1561/0700000027

Barros, P. & Braun, G. (2017). Upcoding in a national health service: the evidence from portugal. *Health Economics*, 26(5), 600–618. https://doi.org/10.1002/hec.3335

Batty, M. & Ippolito, B. (2017). Financial incentives, hospital care, and health outcomes: Evidence from fair pricing laws. *American Economic Journal: Economic Policy*, 9(2), 28–56. https://doi.org/10.1257/pol.20160060

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71, 117–120. https://doi.org/10.1016/j.euroecorev.2014.07.003

Bohnet, I. & Huck, S. (2004). Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American Economic Review*, 94(2), 362–366. https://doi.org/10.1257/0002828041301506

Bolton, G. E., Katok, E., & Ockenfels, A. (2004). How effective are electronic reputation mechanisms? an experimental investigation. *Management Science*, 50(11), 1587–1602. https://doi.org/10.1287/mnsc.1030.0199

Brosig-Koch, J., Hehenkamp, B., & Kokot, J. (2017a). The effects of competition on medical service provision. *Health Economics*, 26, 6–20. https://doi.org/10.1002/hec.3583

Brosig-Koch, J., Hennig-Schmidt, H., Kairies, N., & Wiesen, D. (2013). How effective are pay-for-performance incentives for physicians? - a laboratory experiment. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2278863

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., & Wiesen, D. (2016). Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision. *Journal of Economic Behavior & Organization*, 131, 17–23. https://doi.org/10.1016/j.jebo.2015.04.011

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., & Wiesen, D. (2017b). The effects of introducing mixed payment systems for physicians: Experimental evidence. *Health Economics*, 26(2), 243–262. https://doi.org/10.1002/hec.3292

Brosig-Koch, J., Kairies-Schwarz, N., & Kokot, J. (2017c). Sorting into payment schemes and medical treatment: A laboratory experiment. *Health Economics*, 26(S3), 52–65. https://doi.org/https://doi.org/10.1002/hec.3616

Brown, D. L. & Clement, F. (2018). Calculating health care waste in washington state: first, do no harm. *JAMA Internal Medicine*, 178(9), 1262–1263. https://doi.org/10.1001/jamainternmed.2018.3516

Cabral, L. & Hortaçsu, A. (2010). The dynamics of seller reputation: evidence from ebay*. *The Journal of Industrial Economics*, 58(1), 54–78. https://doi.org/https://doi.org/10.1111/j.1467-6451.2010.00405.x

Cai, H., Jin, G. Z., Liu, C., & Zhou, L.-A. (2014). Seller reputation: From word-of-mouth to centralized feedback. *International Journal of Industrial Organization*, 34, 51–65. https://doi.org/10.1016/j.ijindorg.2014.03.002

Chao, M. & Larkin, I. (2022). Regulating conflicts of interest in medicine through public disclosure: Evidence from a physician payments sunshine law. *Management Science*, 68(2), 1078–1094. https://doi.org/10.1287/mnsc.2020.3940

Chen, A. & Goldman, D. (2016). Health care spending: Historical trends and new directions. *Annual Review of Economics*, 8(1), 291–319. https://doi.org/10.1146/annurev-economics-080315-015317

Chen, P.-Y., Hong, Y., & Liu, Y. (2018). The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Science*, 64(10), 4629–4647. https://doi.org/10.1287/mnsc.2017.2852

Clemens, J. & Gottlieb, J. D. (2014). Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review*, 104(4), 1320–1349. https://doi.org/10.1257/aer.104.4.1320

Cook, A. & Averett, S. (2020). Do hospitals respond to changing incentive structures? evidence from medicare's 2007 drg restructuring. *Journal of Health Economics*, 73, 102319. https://doi.org/https://doi.org/10.1016/j.jhealeco.2020.102319

Currie, J., Lin, W., & Meng, J. (2014). Addressing antibiotic abuse in china: An experimental audit study. *Journal of development economics*, 110, 39–51. https://doi.org/10.1016/j.jdeveco.2014.05.006

Currie, J., Lin, W., & Zhang, W. (2011). Patient knowledge and antibiotic abuse: Evidence from an audit study in china. *Journal of Health Economics*, 30(5), 933–949. https://doi.org/10.1016/j.jhealeco.2011.05.009

Cutler, D. M. (1995). The incidence of adverse medical outcomes under prospective payment. *Econometrica*, 63(1), 29–50. https://doi.org/10.2307/2951696

Dafny, L. S. (2005). How do hospitals respond to price changes? *American Economic Review*, 95(5), 1525–1547. https://doi.org/10.1257/000282805775014236

Dai, T., Akan, M., & Tayur, S. (2017). Imaging room and beyond: The underlying economics behind physicians' test-ordering behavior in outpatient services. *Manufacturing & Service Operations Management*, 19(1), 99–113. https://doi.org/10.1287/msom.2016.0594

Darby, M. R. & Karni, E. (1973). Free competition and the optimal amount of fraud. *The Journal of Law & Economics*, 16(1), 67–88. http://www.jstor.org/stable/724826

Das, J. & Hammer, J. (2007). Money for nothing: The dire straits of medical practice in delhi, india. *Journal of Development Economics*, 83(1), 1–36. https://doi.org/https://doi.org/10.1016/j.jdeveco.2006.05.004

Das, J., Holla, A., Mohpal, A., & Muralidharan, K. (2016). Quality and accountability in health care delivery: Audit-study evidence from primary care in india. *American Economic Review*, 106(12), 3765–99. https://doi.org/10.1257/aer.20151138

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10), 1407–1424.

Dellarocas, C. (2005). Reputation mechanism design in online trading environments with pure moral hazard. *Information Systems Research*, 16(2), 209–230.

Dellarocas, C. (2006). Reputation mechanisms. *Handbook on Economics and Information Systems*, 1–38. https://doi.org/10.1287/isre.1050.0054

Doing-Harris, K., Mowery, D. L., Daniels, C., Chapman, W. W., & Conway, M. (2016). *Understanding patient satisfaction with received healthcare services: A natural language processing approach*. http://europepmc.org/abstract/MED/28269848

Domenighetti, G., Casabianca, A., Gutzwiller, F., & Martinoli, S. (1993). Revisiting the most informed consumer of surgical services: The physician-patient. *International Journal of Technology Assessment in Health Care*, 9(4), 505–513. https://doi.org/10.1017/S0266462300005420

Doyle, J. J. (2005). Health insurance, treatment and outcomes: Using auto accidents as health shocks. *The Review of Economics and Statistics*, 87(2), 256–270. https://doi.org/10.1162/0034653053970348

Dulleck, U. & Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, 44(1), 5–42. https://doi.org/10.1257/002205106776162717

Dulleck, U., Kerschbamer, R., & Sutter, M. (2011). The economics of credence goods: An experiment on the role of liability, verifiability, reputation, and competition. *The American Economic Review*, 101(2), 526–555. http://www.jstor.org/stable/29783682

Dunn, A. & Shapiro, A. H. (2014). Do physicians possess market power? *The Journal of Law and Economics*, 57(1), 159–193. https://doi.org/10.1086/674407

Ely, J., Fudenberg, D., & Levine, D. K. (2008). When is reputation bad? *Games and Economic Behavior*, 63(2), 498–526. https://doi.org/10.1016/j.geb.2006.08.007

Ely, J. C. & Välimäki, J. (2003). Bad reputation. *The Quarterly Journal of Economics*, 118(3), 785–814. http://www.jstor.org/stable/25053923

Fang, L. (2022). The effects of online review platforms on restaurant revenue, consumer learning, and welfare. *Management Science*, 68(11), 8116–8143. https://doi.org/10.1287/mnsc.2021.4279

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171–178. https://doi.org/10.1007/s10683-006-9159-4

Galizzi, M. M. & Wiesen, D. (2018). Behavioral experiments in health economics. *Oxford Research Encyclopedia of Economics and Finance*. https://doi.org/10.1093/acrefore/9780190625979.013.244

Geruso, M. & Layton, T. (2019). Upcoding: Evidence from medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3), 984–1026. https://doi.org/10.1086/704756

Godager, G., Hennig-Schmidt, H., & Iversen, T. (2016). Does performance disclosure influence physicians' medical decisions? an experimental study. *Journal of Economic Behavior & Organization*, 131, 36–46. https://doi.org/https://doi.org/10.1016/j.jebo.2015.10.005

Gottschalk, F., Mimra, W., & Waibel, C. (2020). Health services as credence goods: a field experiment. *The Economic Journal*, 130(629), 1346–1383. https://doi.org/10.1093/ej/ueaa024

Green, E. P. (2014). Payment systems in the healthcare industry: An experimental study of physician incentives. *Journal of Economic Behavior & Organization*, 106, 367–378. https://doi.org/https://doi.org/10.1016/j.jebo.2014.05.009

Gruber, J. & Owings, M. (1996). Physician financial incentives and cesarean section delivery. *The RAND Journal of Economics*, 27(1), 99–123. http://www.jstor.org/stable/2555794

Han, J., Kairies-Schwarz, N., & Vomhof, M. (2017). Quality competition and hospital mergers-an experiment. *Health Economics*, 26, 36–51. https://dx.doi.org/10.1002/hec.3574

Han, J., Kairies-Schwarz, N., & Vomhof, M. (2020). Quality provision in competitive health care markets: Individuals vs. teams (no. 839). *Ruhr Economic Papers*. https://doi.org/10.4419/86788972

Hennig-Schmidt, H., Selten, R., & Wiesen, D. (2011). How payment systems affect physicians' provision behaviour—an experimental investigation. *Journal of Health Economics*, 30(4), 637–646. https://doi.org/https://doi.org/10.1016/j.jhealeco.2011.05.001

Hong, Y. A., Liang, C., Radcliff, T. A., Wigfall, L. T., & Street, R. L. (2019). What do patients say about doctors online? a systematic review of studies on patient online reviews. *J Med Internet Res*, 21(4), e12521. https://doi.org/10.2196/12521

Huck, S., Lünser, G., Spitzer, F., & Tyran, J.-R. (2016a). Medical insurance and free choice of physician shape patient overtreatment: A laboratory experiment. *Journal of Economic Behavior & Organization*, 131, 78–105. https://doi.org/https://doi.org/10.1016/j.jebo.2016.06.009

Huck, S., Lünser, G. K., & Tyran, J.-R. (2012). Competition fosters trust. *Games and Economic Behavior*, 76(1), 195–209. https://doi.org/https://doi.org/10.1016/j.geb.2012.06.010

Huck, S., Lünser, G. K., & Tyran, J.-R. (2016b). Price competition and reputation in markets for experience goods: an experimental study. *The RAND Journal of Economics*, 47(1), 99–117. https://doi.org/https://doi.org/10.1111/1756-2171.12120

Iizuka, T. (2007). Experts' agency problems: Evidence from the prescription drug market in japan. *RAND Journal of Economics*, 38(3), 844 – 862. https://doi.org/10.1111/j.0741-6261.2007.00115.x

Januleviciute, J., Askildsen, J. E., Kaarboe, O., Siciliani, L., & Sutton, M. (2016). How do hospitals respond to price changes? evidence from norway. *Health Economics*, 25(5), 620–636. https://doi.org/10.1002/hec.3179

Jürges, H. & Köberlein, J. (2015). What explains drg upcoding in neonatology? the roles of financial incentives and infant health. *Journal of Health Economics*, 43, 13–26. https://doi.org/https://doi.org/10.1016/j.jhealeco.2015.06.001

Kairies, N. & Krieger, M. (2013). How do non-monetary performance incentives for physicians affect the quality of medical care? - a laboratory experiment. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2278866

Kairies-Schwarz, N., Kokot, J., Vomhof, M., & Weßling, J. (2017). Health insurance choice and risk preferences under cumulative prospect theory – an experiment. *Journal of Economic Behavior & Organization*, 137, 374–397. https://doi.org/https://doi.org/10.1016/j.jebo.2017.03.012

Kerschbamer, R., Neururer, D., & Sutter, M. (2016). Insurance coverage of customers induces dishonesty of sellers in markets for credence goods. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27), 7454–7458. https://doi.org/10.1073/pnas.1518015113

Kerschbamer, R., Neururer, D., & Sutter, M. (2023). Credence goods markets, online information and repair prices: A natural field experiment. *Journal of Public Economics*, 222, 104891. https://doi.org/https://doi.org/10.1016/j.jpubeco.2023.104891

Kerschbamer, R., Sutter, M., & Dulleck, U. (2017). How social preferences shape incentives in (experimental) markets for credence goods. *The Economic Journal*, 127(600), 393–416. https://doi.org/https://doi.org/10.1111/ecoj.12284

Kesternich, I., Schumacher, H., & Winter, J. (2015). Professional norms and physician behavior: Homo oeconomicus or homo hippocraticus ? *Journal of Public Economics*, 131, 1–11. https://doi.org/10.1016/j.jpubeco.2015.08.009

Klein, B. & Leffler, K. B. (1981). The role of market forces in assuring contractual performance. *Journal of Political Economy*, 89(4), 615–641.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2), 245–252. https://doi.org/https://doi.org/10.1016/0022-0531(82)90029-1

Lagarde, M. & Blaauw, D. (2017). Physicians' responses to financial and social incentives: A medically framed real effort experiment. *Social Science & Medicine*, 179, 147–159. https://doi.org/https://doi.org/10.1016/j.socscimed.2017.03.002

Lim, T. O., Sorays, A., Ding, L. M., & Morad, Z. (2002). Assessing doctors' competence: application of cusum technique in monitoring doctors' performance. *International Journal for Quality in Health Care*, 14(3), 251–258. https://doi.org/10.1093/oxfordjournals.intqhc.a002616

Liu, M., Brynjolfsson, E., & Dowlatabadi, J. (2021). Do digital platforms reduce moral hazard? the case of uber and taxis. *Management Science*. https://doi.org/10.1287/mnsc.2020.3721

Lu, S. F. & Rui, H. (2018). Can we trust online physician ratings? evidence from cardiac surgeons in florida. *Management Science*, 64(6), 2557–2573. https://doi.org/10.1287/mnsc.2017.2741

López, A., Detz, A., Ratanawongsa, N., & Sarkar, U. (2012). What patients say about their doctors online: A qualitative content analysis. *Journal of General Internal Medicine*, 27(6), 685–692. https://doi.org/10.1007/s11606-011-1958-4

Mimra, W., Rasch, A., & Waibel, C. (2016). Price competition and reputation in credence goods markets: Experimental evidence. *Games and Economic Behavior*, 100, 337–352. https://doi.org/10.1016/j.geb.2016.09.012

Moreno, A. & Terwiesch, C. (2014). Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research*, 25(4), 865–886.

OECD (2023). *Health at a Glance 2023*. https://doi.org/https://doi.org/10.1787/7a7afb35-en

Okike, K., Peter-Bibb, T. K., Xie, K. C., & Okike, O. N. (2016). Association between physician online rating and quality of care. *Journal of medical Internet research*, 18(12), e324.

Parkinson, B., Meacock, R., & Sutton, M. (2019). How do hospitals respond to price changes in emergency departments? *Health Economics*, 28(7), 830–842. https://doi.org/10.1002/hec.3890

Pasero, C. & McCaffery, M. (2001). The undertreatment of pain: Are providers accountable for it? *AJN The American Journal of Nursing*, 101(11). https://journals.lww.com/ajnonline/Fulltext/2001/11000/The_Undertreatment_of_Pain

Rajgopal, S. & White, R. (2019). Cheating when in the hole: The case of new york city taxis. *Accounting, Organizations and Society*, 79, 101070. https://doi.org/10.1016/j.aos.2019.101070

Reif, S., Hafner, L., & Seebauer, M. (2020). Physician behavior under prospective payment schemes—evidence from artefactual field and lab experiments. *International Journal of Environmental Research and Public Health*, 17(15), 5540. https://doi.org/10.3390/ijerph17155540

Resnick, P. & Zeckhauser, R. (2002). *Trust among strangers in internet transactions: Empirical analysis of eBay' s reputation system*, volume 11 of *Advances in Applied Microeconomics*, 127–157. Emerald Group Publishing Limited. https://doi.org/10.1016/S0278-0984(02)11030-3

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2), 79–101. https://doi.org/10.1007/s10683-006-4309-2

Rice, S. C. (2012). Reputation and uncertainty in online markets: An experimental study. *Information Systems Research*, 23(2), 436–452. https://doi.org/10.1287/isre.1110.0362

Saifee, D. H., Bardhan, I. R., Lahiri, A., & Zheng, Z. (2019). Adherence to clinical guidelines, electronic health record use, and online reviews. *Journal of Management Information Systems*, 36(4), 1071–1104.

Saifee, D. H., Zheng, Z. E., Bardhan, I. R., & Lahiri, A. (2020). Are online reviews of physicians reliable indicators of clinical outcomes? a focus on chronic disease management. *Information Systems Research*, 31(4), 1282–1300. https://doi.org/10.1287/isre.2020.0945

Schneider, H. S. (2012). Agency problems and reputation in expert services: Evidence from auto repair. *The Journal of Industrial Economics*, 60(3), 406–433. https://doi.org/https://doi.org/10.1111/j.1467-6451.2012.00485.x

Shapiro, C. (1982). Consumer information, product quality, and seller reputation. *The Bell Journal of Economics*, 13(1), 20–35.

Shigeoka, H. & Fushimi, K. (2014). Supplier-induced demand for newborn treatment: Evidence from japan. *Journal of Health Economics*, 35, 162–178. https://doi.org/10.1016/j.jhealeco.2014.03.003

Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8, 321–340. https://doi.org/10.1146/annurev-economics-080315-015325

Xu, Y., Armony, M., & Ghose, A. (2021). The interplay between online reviews and physician demand: An empirical investigation. *Management Science*, 67(12), 7344–7361. https://doi.org/10.1287/mnsc.2020.3879

Yelp (2020). *Fast facts*. https://www.yelp-press.com/company/fast-facts/default.aspx

Appendix A Results for healthcare and neutral frame

A.1 Main conditions healthcare and neutral frame

In the following, we present the results for the main conditions, for both the healthcare and neutral frame separately.

The healthcare frame sessions were conducted in the laboratory for experimental economics research at the University of Innsbruck. A total of 192 students participated, 48 subjects in each condition. Subjects earned 25.40 euros on average and sessions lasted approximately 105 minutes. Table A1 provides an overview of the sample characteristics.

The neutral frame sessions took place in the laboratory for experimental economic research (experimenTUM) at the Technical University of Munich (TUM School of Management). A total of 184 students participated, 48 subjects in each condition except for condition (***Exp+Rating-N***), including only 40 subjects due to recruitment challenges. Participants in the lab in Munich are on average 2.9 years older than those in Innsbruck. Subjects earned 30.70 euros on average and sessions lasted approximately 105 minutes. Table A2 provides an overview of the sample characteristics.

In the following, all tables and figures reported in Section 4 are presented for the healthcare and neutral frame separately. Subsequently, Section A.2 provides a comparison of results in condition ***Rating-N*** between participants in the laboratory of experimental economics at the University of Innsbruck and the Technical University of Munich.

A.1.1 Tables for conditions healthcare and neutral frame

Table A1: Descriptive statistics (healthcare frame)

	Markets without personal experience		Market with personal experience	
	Baseline-H	*Rating-H*	*Experience-H*	*Exp+Rating-H*
Male (in %)	41.7	52.1	43.8	54.2
Age (in years)	22.3	22.8	22.8	21.3
	(2.8)	(3.1)	(4.1)	(1.9)
Relative School Performance	72.2	70.9	70.4	68.9
	(18.0)	(22.2)	(20.1)	(19.3)
Number of Physician Visits last year	4.9	5.8	4.9	4.6
	(4.8)	(6.6)	(5.2)	(4.1)
Exp. with incorrect physician behavior (in %)	56.3	45.8	27.1	52.1
	(50.1)	(50.4)	(44.9)	(50.5)
Exp. with physician recommendation (in %)	81.3	75.0	77.1	85.4
	(39.4)	(43.8)	(42.5)	(35.7)
Business/Economics major (in %)	29.2	52.1	33.3	45.8
	(45.9)	(50.5)	(47.6)	(50.4)
Frequency of practicing Religion (in %)				
Never	54.2	58.3	77.1	39.6
Rarely	33.3	33.3	18.8	39.6
Often	12.5	8.3	4.2	20.8
Extraversion	3.5	3.4	3.3	3.7
	(1.0)	(1.0)	(0.9)	(0.9)
Agreeableness	3.3	3.1	3.2	3.5
	(0.8)	(1.0)	(1.0)	(0.8)
Conscientiousness	3.6	3.4	3.3	3.5
	(0.8)	(0.8)	(0.9)	(0.8)
Neuroticism	2.9	3.0	2.9	2.8
	(0.9)	(0.8)	(0.9)	(0.9)
Openness	3.8	3.5	3.5	3.8
	(0.9)	(0.9)	(1.2)	(1.0)
Amount donated to charity in a DG (in euros)	3.2	4.0	4.9	4.1
	(3.2)	(3.4)	(4.2)	(3.5)
Risk Aversion	11.7	12.5	12.3	11.1
	(2.8)	(3.9)	(3.0)	(3.4)
Trustworthiness	0.3	0.3	0.4	0.5
	(0.3)	(0.2)	(0.3)	(0.2)
Liar (in %)	91.7	83.3	85.4	81.3
	(27.9)	(37.7)	(35.7)	(39.4)
Experimental payoff (physicians)	75.3	42.3	40.3	38.0
	(23.5)	(18.9)	(21.5)	(20.1)
Experimental payoff (patients)	-20.3	48.1	48.8	54.0
	(30.3)	(16.3)	(18.1)	(15.0)

Note: We analyze six independent markets in every experimental condition with a health frame. In each market, four consumers and four experts interact. Means (standard deviations). Background variables were measured in additional experiments and a post-experimental questionnaire (see Appendix F for a detailed description of all these measures). Amount donated to a charity in a dictator game: donation of up to 12 euros as a measure of altruism. Risk aversion: number of safe choices in a choice list with 20 binary decision problems between a risky prospect and a safe option. Trustworthiness: share sent back to the first-mover in a trust game. Liar: dummy variable = 1 if someone reports 4 or more correct dice rolls out of 12 in a lying task. BIG 5 personality traits (extraversion, agreeableness, conscientiousness, neuroticism, and openness) measured with a 10-item BIG 5 questionnaire. Self-reported relative school performance as a proxy for IQ. The experimental payoff is the sum of payoffs in ECUs generated by participants over the 16 periods (not the payout they received at the session's end).

Table A2: Descriptive statistics (neutral frame)

	Markets without personal experience		Market with personal experience	
	Baseline-N	***Rating-N***	***Experience-N***	***Exp+Rating-N***
Male (in %)	52.1	54.2	47.9	55.0
Age (in years)	25.9	26.3	24.7	23.7
	(6.3)	(7.1)	(5.8)	(6.9)
Relative School Performance	79.0	79.4	82.3	80.4
	(14.4)	(17.4)	(13.7)	(18.7)
Number of Physician Visits last year	3.8	5.4	5.5	4.8
	(2.7)	(8.7)	(5.6)	(3.6)
Exp. with incorrect physician behavior (in %)	39.6	25.0	35.4	45.0
	(49.4)	(43.8)	(48.3)	(50.4)
Exp. with physician recommendation (in %)	64.6	72.9	58.3	77.5
	(48.3)	(44.9)	(49.8)	(42.3)
Business/Economics major (in %)	43.8	35.4	47.9	42.5
	(50.1)	(48.3)	(50.5)	(50.1)
Frequency of practicing Religion (in %)				
Never	50.0	54.2	54.2	60.0
Rarely	37.5	39.6	29.2	27.5
Often	12.5	6.3	16.7	12.5
Extraversion	3.1	3.2	3.1	3.3
	(0.9)	(1.1)	(1.0)	(1.0)
Agreeableness	3.2	3.2	3.3	3.1
	(0.6)	(0.8)	(0.6)	(0.9)
Conscientiousness	3.4	3.4	3.7	3.8
	(0.8)	(0.7)	(0.7)	(0.9)
Neuroticism	3.1	2.8	2.9	2.6
	(0.8)	(1.0)	(1.2)	(1.0)
Openness	3.6	3.5	3.5	3.1
	(0.9)	(1.0)	(1.0)	(1.0)
Amount donated to charity in a DG (in euros)	3.6	3.7	4.0	4.8
	(3.9)	(4.0)	(3.3)	(3.7)
Risk Aversion	12.8	12.1	12.2	11.6
	(3.3)	(3.1)	(3.3)	(3.8)
Trustworthiness	0.3	0.4	0.4	0.4
	(0.2)	(0.2)	(0.2)	(0.2)
Liar (in %)	81.3	77.1	79.2	65.0
	(39.4)	(42.5)	(41.0)	(48.3)
Experimental payoff (experts)	66.7	38.3	44.2	38.5
	(23.8)	(21.0)	(29.6)	(21.4)
Experimental payoff (consumers)	7.8	54.7	44.9	46.9
	(32.0)	(20.3)	(13.6)	(16.5)

Note: We analyze six independent markets in every experimental condition with a neutral frame except for ***Exp+Rating-N***, where we only have five markets. In each market, four consumers and four experts interact. Means (standard deviations). Background variables were measured in additional experiments and a post-experimental questionnaire (see Appendix F for a detailed description of all these measures). Amount donated to a charity in a dictator game: donation of up to 12 euros as a measure of altruism. Risk aversion: number of safe choices in a choice list with 20 binary decision problems between a risky prospect and a safe option. Trustworthiness: share sent back to the first-mover in a trust game. Liar: dummy variable = 1 if someone reports 4 or more correct dice rolls out of 12 in a lying task. BIG 5 personality traits (extraversion, agreeableness, conscientiousness, neuroticism, and openness) measured with a 10-item BIG 5 questionnaire. Self-reported relative school performance as a proxy for IQ. The experimental payoff is the sum of payoffs in ECUs generated by participants over the 16 periods (not the payout they received at the session's end).

Table A3: Overview of results (means - healthcare frame).

	Markets without personal experience		Markets with personal experience		*p-values of MWU*[1]		
	Baseline-H	***Rating-H***	***Experience-H***	***Exp+Rating-H***	***Baseline-H vs Rating-H***	***Rating-H vs Experience-H***	***Experience-H vs Exp+Rating-H***
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Expert behavior							
Undertreatment (in %)	64.72	5.81	6.81	4.62	0.002	0.794	0.900
Overcharging (in %)	92.03	47.94	36.89	43.64	0.002	0.093	0.554
Overtreatment (in %)	0.00	4.30	6.47	2.16	0.455	0.546	0.546
Consumer decisions							
Interaction (in %)	93.75	98.96	99.74	99.21	1.000	0.424	0.546
Feedback (in %)	-	93.62	-	90.58			
Star-rating	-	3.66	-	3.73			
Market outcomes							
Efficiency (in %)	70.70	96.21	95.42	95.94	0.002	0.849	0.974
Consumer Surplus (in ECUs)	-1.27	3.01	3.05	3.37	0.002	0.937	0.589
Observations	48	48	48	48			

Note: Results from conditions with a healthcare market frame. We analyze six independent markets in every experimental condition. In each market, four consumers and four experts interact. The experimental conditions are: ***Baseline-H***, ***Experience-H***, ***Rating-H***, and ***Exp+Rating-H***. Please refer to Section 3.2 for a description of the experimental conditions. See Table 3 for a description of the outcome variables.

[1] Mann-Whitney U-tests for pairwise differences between conditions with matching groups of 8 subjects as one independent observation. p-values are adjusted for the small sample size, using Fisher's exact test.

Table A4: Overview of results (means - neutral frame)

	Markets without personal experience		Markets with personal experience		*p-values of MWU*[1]		
	Baseline-N	***Rating-N***	***Experience-N***	***Exp+Rating-N***	***Baseline-N vs Rating-N***	***Rating-N vs Experience-N***	***Experience-N vs Exp+Rating-N***
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Expert behavior							
Undertreatment (in %)	40.13	8.61	11.63	14.43	0.058	0.368	0.273
Overcharging (in %)	79.64	39.86	48.35	34.90	0.015	0.563	0.165
Overtreatment (in %)	2.10	0.91	1.65	4.35	0.697	0.848	0.545
Consumer decisions							
Interaction (in %)	98.18	100.00	100.00	97.50	0.455	1.000	0.030
Feedback (in %)	-	96.35	-	91.05			
Star-rating	-	3.88	-	3.55			
Market outcomes							
Efficiency (in %)	82.90	96.90	95.64	92.56	0.009	0.485	0.247
Consumer Surplus (in ECUs)	0.48	3.420	2.81	2.93	0.015	0.180	1.000
Observations	48	48	48	40			

Note: Results from conditions with a neutral market frame. We analyze six independent markets in every experimental condition except for ***Exp+Rating-N***, where we only have five markets. In each market, four consumers and four experts interact. The experimental conditions are: ***Baseline-N***, ***Rating-N***, ***Experience-N***, and ***Exp+Rating-N***. Please refer to Section 3.2 for a description of the main experimental conditions. See Table 3 for a description of the outcome variables.

[1] Mann-Whitney U-tests for pairwise differences between conditions with matching groups of 8 subjects as one independent observation. p-values are adjusted for the small sample size, using Fisher's exact test.

Table A5: Average treatment effects (healthcare frame)

	Undertreatment		**Overcharging**		**Efficiency**	**Consumer Surplus**
	(1)	(2)	(3)	(4)	(5)	(6)
Levels in ***Baseline-H***	0.690	0.647	0.943	0.935	0.707	-1.266
	(0.171)	(0.093)	(0.041)	(0.035)	(0.068)	(0.642)
	Marginal Treatment Effects					
Rating-H	-0.584***	-0.553***	-0.397***	-0.414***	0.255***	4.271***
	(0.164)	(0.121)	(0.052)	(0.050)	(0.069)	(0.686)
Experience-H	-0.585***	-0.539***	-0.491***	-0.428***	0.253***	4.318***
	(0.170)	(0.134)	(0.057)	(0.072)	(0.070)	(0.718)
Exp+Rating-H	-0.615***	-0.544***	-0.492***	-0.455***	0.264***	4.638***
	(0.168)	(0.125)	(0.070)	(0.063)	(0.068)	(0.689)
Period	+***	+***	+***	+***	-***	-***
	Additional Games					
Amount donated to charity		-***		-*		
Liar (yes)		not sig.		not sig.		
Trustworthiness		not sig.		not sig.		
Covariates		✓		✓		
	p-values from post-estimation Wald-Test					
Rating-H vs ***Experience-H***	*0.985*	*0.737*	*0.060*	*0.800*	*0.894*	*0.907*
Rating-H vs ***Exp+Rating-H***	*0.300*	*0.830*	*0.146*	*0.447*	*0.402*	*0.290*
Experience-H vs ***Exp+Rating-H***	*0.474*	*0.924*	*0.987*	*0.724*	*0.461*	*0.429*
Observations	766		738		1536	1536
Number of Groups	24		24		24	24

Note: The table presents results from multilevel models with random effects at the market and individual levels (undertreatment & overcharging: columns 1-4) or at the market level for market efficiency (column 5) and consumer surplus (column 6) from conditions with a healthcare market frame. See Table 3 for a description of the outcome variables. We report effects as marginal effects, calculated as differences in the expected probabilities between the experimental condition in question and the baseline condition. Please refer to Section 3.2 for a description of the experimental conditions. All regressions include time trends. **Covariates**: Gender, age, BIG 5 personality traits (extraversion, agreeableness, conscientiousness, neuroticism, openness) measured with a 10-item BIG 5 questionnaire, whether the participant is a business/economics major, self-reported frequency of practicing religion, number of physician visits in the past 12 months, an indicator for experience with incorrect physician behavior, an indicator for experience with physician recommendations, relative school performance as a proxy for IQ, a measure for altruism (the amount donated to charity in a dictator game), an indicator whether the participant is classified as a liar (if reporting 4 or more correct dice rolls out of 12 in a lying task), and trustworthiness measured in a standard trust game. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Average treatment effects (neutral frame)

	Undertreatment		Overcharging		Efficiency	Consumer Surplus
	(1)	(2)	(3)	(4)	(5)	(6)
Levels in ***Baseline-N***	0.379	0.389	0.806	0.797	0.829	0.484
	(0.078)	(0.075)	(0.071)	(0.061)	(0.045)	(0.728)
	Marginal Treatment Effects					
Rating-N	-0.240***	-0.274***	-0.328***	-0.373***	0.140***	2.935***
	(0.081)	(0.078)	(0.101)	(0.102)	(0.046)	(0.811)
Experience-N	-0.189**	-0.190**	-0.270***	-0.326***	0.127***	2.323***
	(0.085)	(0.080)	(0.097)	(0.086)	(0.047)	(0.751)
Exp+Rating-N	-0.132	-0.162**	-0.405***	-0.426***	0.097*	2.444***
	(0.086)	(0.073)	(0.094)	(0.096)	(0.050)	(0.778)
Period	+***	+***	+***	+***	-***	-***
	Additional Games					
Amount donated to charity		not sig.		not sig.		
Liar (yes)		not sig.		not sig.		
Trustworthiness		-*		not sig.		
Covariates		✓		✓		
	p-values from post-estimation Wald-Test					
Rating-N vs ***Experience-N***	*0.186*	*0.170*	*0.531*	*0.596*	*0.402*	*0.129*
Rating-N vs ***Exp+Rating-N***	*0.010*	*0.037*	*0.402*	*0.613*	*0.056*	*0.276*
Experience-N vs ***Exp+Rating-N***	*0.264*	*0.700*	*0.119*	*0.290*	*0.187*	*0.716*
Observations	709		748		1472	1472
Number of Groups	23		23		23	23

Note: The table presents results from multilevel models with random effects at the market and individual levels (undertreatment & overcharging: columns 1-4) or at the market level for market efficiency (column 5) and consumer surplus (column 6) from conditions with a neutral market frame. See Table 3 for a description of the outcome variables. We report effects as marginal effects, calculated as differences in the expected probabilities between the experimental condition in question and the baseline condition. Please refer to Section 3.2 for a description of the experimental conditions. All regressions include time trends. **Covariates**: Gender, age, BIG 5 personality traits (extraversion, agreeableness, conscientiousness, neuroticism, openness) measured with a 10-item BIG 5 questionnaire, whether the participant is a business/economics major, self-reported frequency of practicing religion, number of physician visits in the past 12 months, an indicator for experience with incorrect physician behavior, an indicator for experience with physician recommendations, relative school performance as a proxy for IQ, a measure for altruism (the amount donated to charity in a dictator game), an indicator whether the participant is classified as a liar (if reporting 4 or more correct dice rolls out of 12 in a lying task), and trustworthiness measured in a standard trust game. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Ratings response (healthcare frame)

	Star rating
Predicted star-rating if patient payoff is 7 ECUs	4.79
	Marginal effects if ...
... payoff is 2 ECUs	-1.490***
	(0.163)
... payoff is -8 ECUs	-4.696***
	(0.143)
Observations	701
Number of groups	12

Note: For this analysis, only the treatments ***Rating-H*** and ***Exp+Rating-H*** are considered. The table presents the marginal treatment effects of multilevel models with random effects at the market and individual levels. The dependent variables are star ratings following an interaction with an expert. Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Ratings response (neutral frame)

	Star rating
Predicted star-rating if patient payoff is 7 ECUs	4.73
	Marginal effects if ...
... payoff is 2 ECUs	-1.346***
	(0.151)
... payoff is -8 ECUs	-4.167***
	(0.362)
Observations	654
Number of groups	11

Note: For this analysis, only the treatments ***Rating-N*** and ***Exp+Rating-N*** are considered. The table presents the marginal treatment effects of multilevel models with random effects at the market and individual levels. The dependent variables are star ratings following an interaction with an expert. Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Associations between payoffs of consumers and their decision to change the expert (health-care frame)

	Change expert
Frequency of change if consumer-profit is 7 ECUs	0.16
	Marginal effects if ...
... patient-profit is 2 ECUs	0.290***
	(0.048)
... patient-profit is -8 ECUs	0.639***
	(0.069)
Observations	720
Number of groups	12

Note: For this analysis, only the treatments ***Experience-H*** and ***Exp+Rating-H*** are considered. The table presents results from a three-level model with random effects at the market and individual levels. The dependent variable is a binary indicator of whether a consumer changed the expert. We report effects as marginal effects, calculated as differences in the expected probabilities between the payoff in question and the maximum profit of 7 ECUs. Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Associations between payoffs of consumers and their decision to change the expert (neutral frame)

	Change expert
Frequency of change if consumer-profit is 7 ECUs	0.34
	Marginal effects if ...
... patient-profit is 2 ECUs	0.185***
	(0.037)
... patient-profit is -8 ECUs	0.474***
	(0.075)
Observations	651
Number of groups	11

Note: For this analysis, only the treatment ***Experience-N*** and ***Exp+Rating-N*** are considered. The table presents results from a three-level model with random effects at the market and individual levels. The dependent variable is a binary indicator of whether a consumer changed the expert. We report effects as marginal effects, calculated as differences in the expected probabilities between the payoff in question and the maximum profit of 7 ECUs. Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Expert visits according to rank (healthcare frame)

	Public and private best	***Public best (private not)***	***Private best (public not)***	***Other***
	(1)	(2)	(3)	(4)
Levels in ***Rating-H***	0.415	0.197	0.116	0.271
	Marginal treatment effects			
Exp+Rating-H	-0.079*	-0.032	0.147***	-0.036
	(0.040)	(0.031)	(0.032)	(0.036)
Experience-H	0.055	-0.155***	0.166***	-0.066*
	(0.042)	(0.026)	(0.033)	(0.036)
Exp+Rating-Priv-H	0.102**	-0.173***	0.123***	-0.052
	(0.041)	(0.025)	(0.032)	(0.036)
	p-values from post-estimation Wald-Tests			
Exp+Rating-H vs ***Experience-H***	0.001	0.000	0.608	0.394
Exp+Rating-H vs ***Exp+Rating-Priv-H***	0.000	0.000	0.515	0.641
Experience-H vs ***Exp+Rating-Priv-H***	0.259	0.240	0.244	0.699

Note: The table presents results from a multinomial logistic regression. We report the predicted frequencies of choosing experts based on public and private ranks in ***Rating-H*** and the difference between ***Rating-H*** and other experimental conditions as marginal effects. E.g. patients in ***Rating-H*** choose the private (but not public) best-rated expert in 11.6% of cases (column 3), while patients in ***Exp+Rating-H*** do so significantly more often, in 26.3% of the cases.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Expert visits according to rank (neutral frame)

	Public and private best	***Public best (private not)***	***Private best (public not)***	***Other***
	(1)	(2)	(3)	(4)
Levels in ***Rating-N***	0.503	0.094	0.139	0.264
	Marginal treatment effects			
Experience-N	-0.222***	-0.066***	0.285***	0.003
	(0.040)	(0.020)	(0.036)	(0.037)
Exp+Rating-N	-0.210***	-0.057***	0.159***	0.005
	(0.042)	(0.029)	(0.036)	(0.039)
	p-values from post-estimation Wald-Tests			
Exp+Rating-N vs ***Experience-N***	0.767	0.000	0.003	0.822

Note: The table presents results from a multinomial logistic regression. We report the predicted frequencies of choosing experts based on public and private ranks in ***Rating-N*** and the difference between ***Rating-N***, ***Experience-N***, and ***Exp+Rating-N*** as marginal effects. E.g. patients in ***Rating-N*** choose the private (but not public) best-rated expert in 13.9% of cases (column 3), while patients in ***Experience-N*** do so significantly more often, in 42.4% of the cases.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.1.2 Figures for conditions healthcare and neutral frame

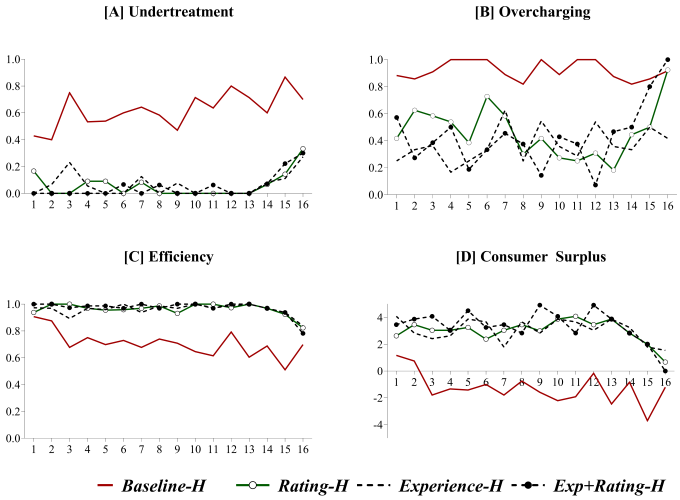


Figure A1: Rate of undertreatment, overcharging, efficiency, and consumer surplus by experimental conditions (healthcare frame).

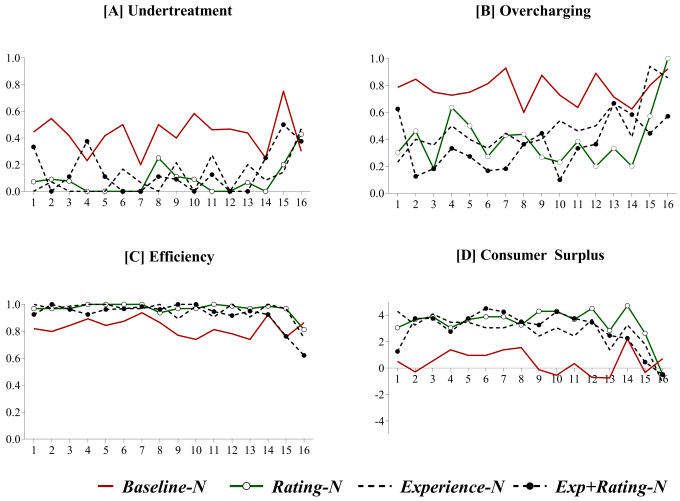


Figure A2: Rate of undertreatment, overcharging, efficiency, and consumer surplus by experimental conditions (neutral frame).

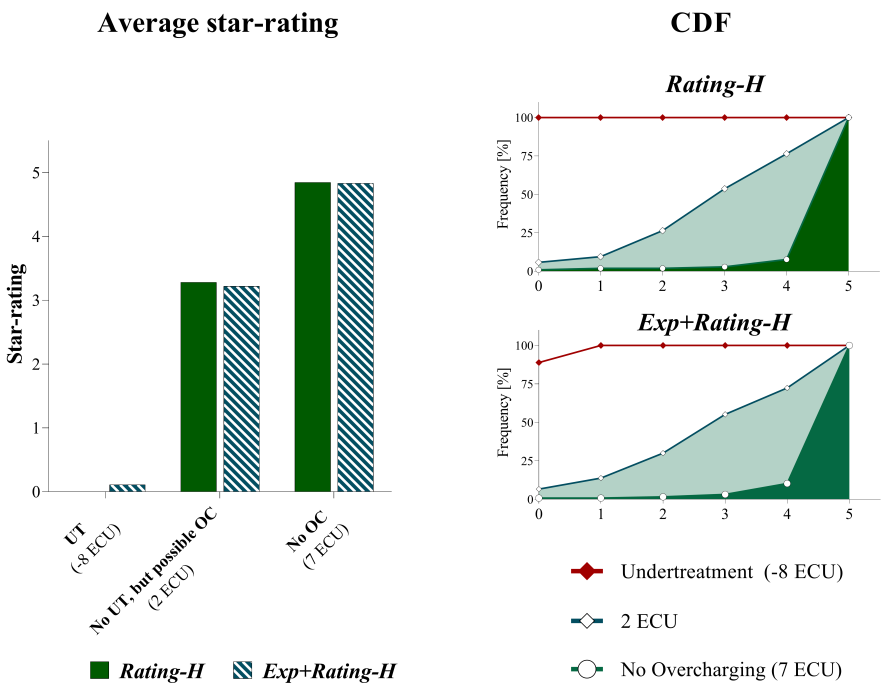


Figure A3: Rating behavior of consumers (healthcare frame). On the left side, we see the mean ratings for each of the possible payoffs of consumers. The right side shows the cumulative distribution function (CDF) of given star ratings, separately for possible payoffs of consumers. If consumers are undertreated, the payoff is -8 ECUs, whereas if they have a minor problem and are treated appropriately, the payoff is 7 ECUs. In the case of a minor problem and appropriate treatment but overcharging, or in the case of a major problem and appropriate treatment with charges, the payoff is 2 ECUs.

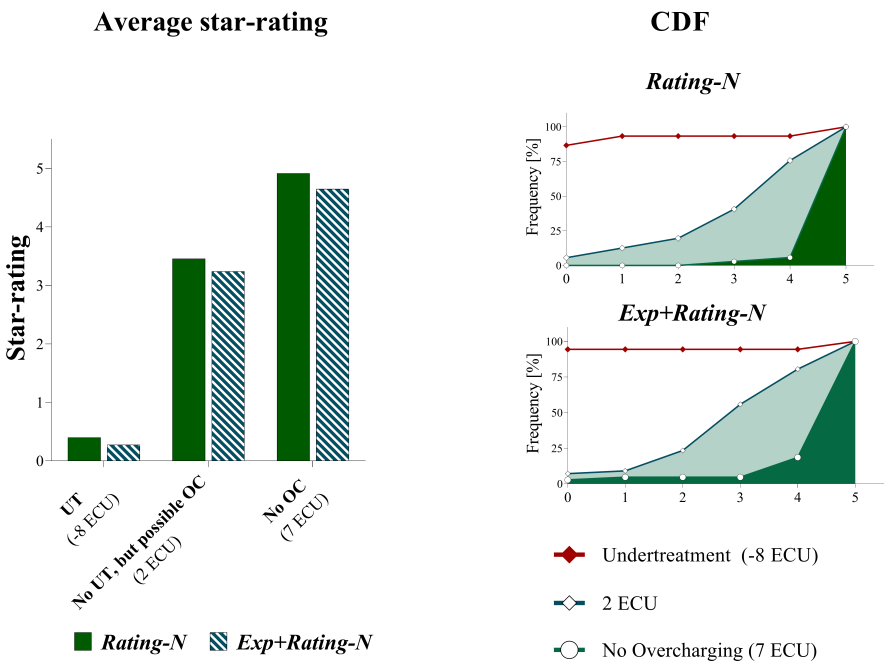


Figure A4: Rating behavior of consumers (neutral frame). On the left side, we see the mean ratings for each of the possible payoffs of consumers. The right side shows the cumulative distribution function (CDF) of given star ratings, separately for possible payoffs of consumers. If consumers are undertreated, the payoff is -8 ECUs, whereas if they have a minor problem and are treated appropriately, the payoff is 7 ECUs. In the case of a minor problem and appropriate treatment but overcharging, or in the case of a major problem and appropriate treatment with charges, the payoff is 2 ECUs.

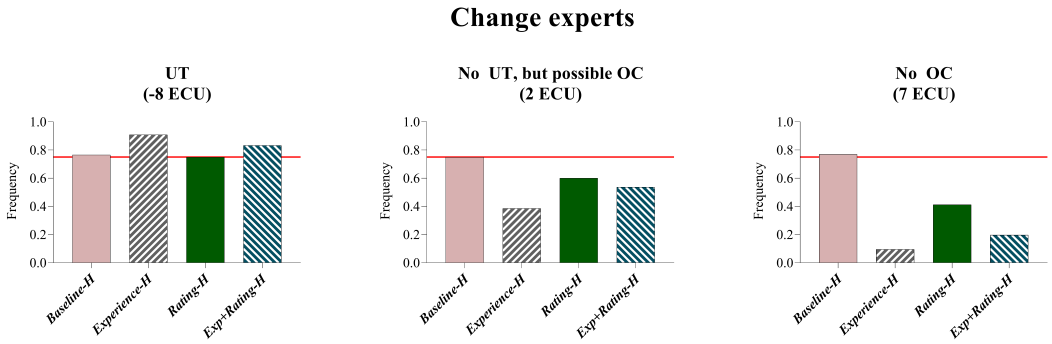


Figure A5: Frequency of a change in expert by realized consumer payoff in a given period (healthcare frame).

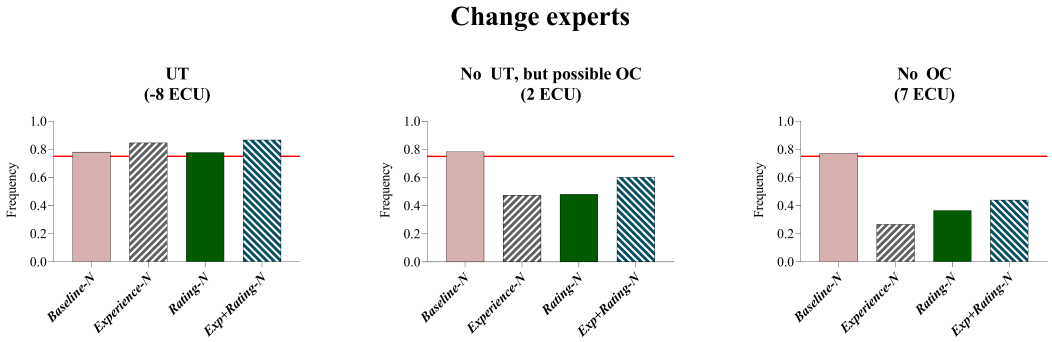


Figure A6: Frequency of a change in expert by realized consumer payoff in a given period (neutral frame).

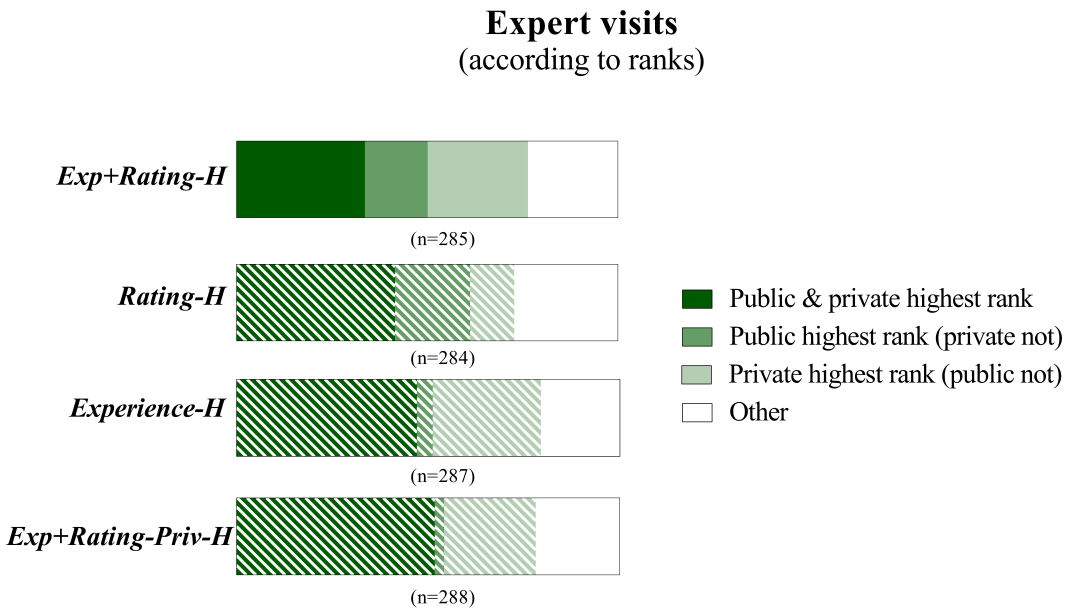


Figure A7: Distribution of consumers' realized expert visits according to relative private and public rankings of experts (healthcare frame). Results for ***Rating-H***, ***Experience-H*** and ***Exp+Rating-Priv-H*** are crosshatched as consumers do not have full information on both private and public rankings in these conditions.

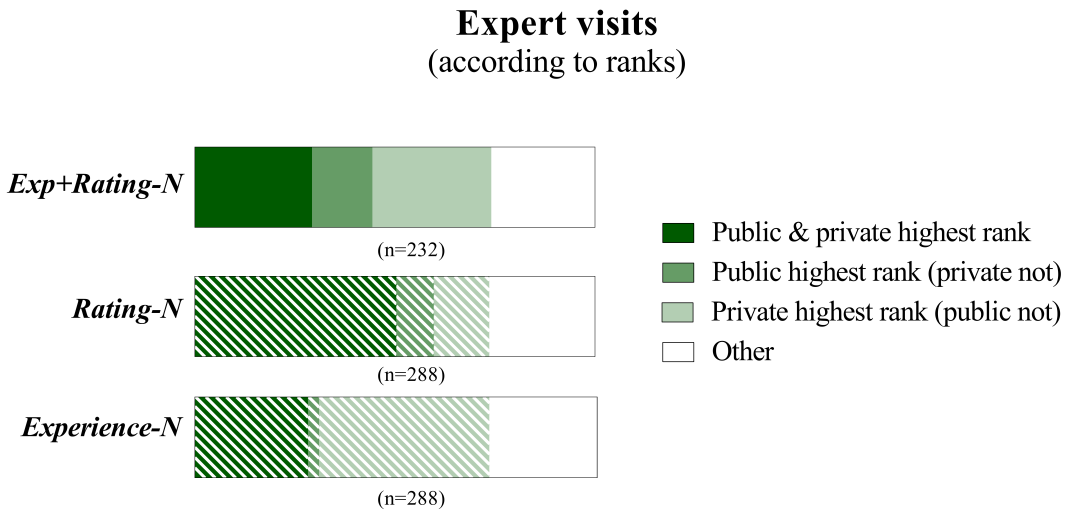


Figure A8: Distribution of consumers' realized expert visits according to relative private and public rankings of experts (neutral frame). Results for ***Rating-N***, and ***Experience-N*** are crosshatched as consumers do not have full information on both private and public rankings in these conditions.

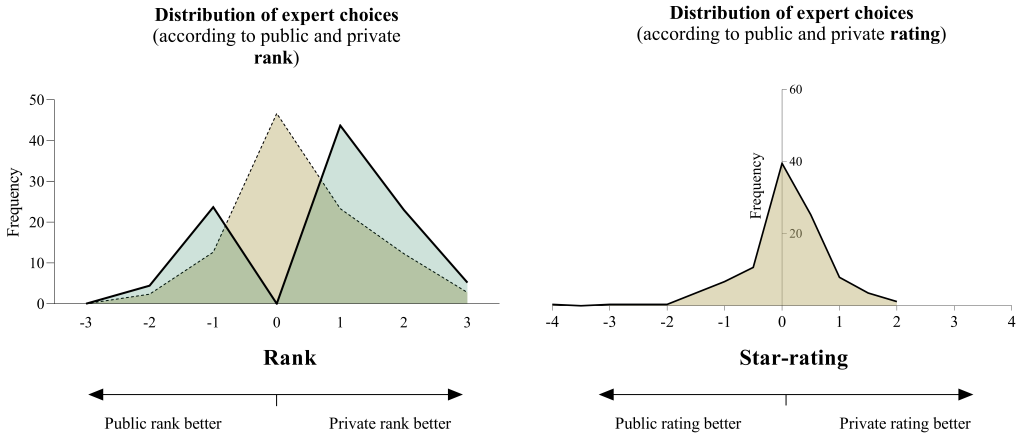


Figure A9: Distribution of selected experts by the spread in public-private rank (left) and rating (right) (healthcare frame). For this figure, only the treatment ***Exp+Rating-H*** is considered. **Left**: The dashed line shows the distribution of choices according to ranks including equal ranks. We observed 253 interactions where consumers chose an expert for whom they had both, private experiences, and a public rating. Of those, consumers selected an expert with equal ranks in 46.6% (118 interactions). The solid line (shaded area) only shows the distribution of selected experts when there was a discrepancy between the private and the public ranking. Testing for normality reveals that the distribution is significantly skewed to the left (-0.458, $p < 0.05$). **Right**: we show the distribution of selected experts according to differences between the private and public average ratings (private average rating - public average rating). Hence, positive (negative) numbers indicate that the expert had a better private (public) rating. The distribution is significantly skewed to the left (-1.032, $p < 0.01$).

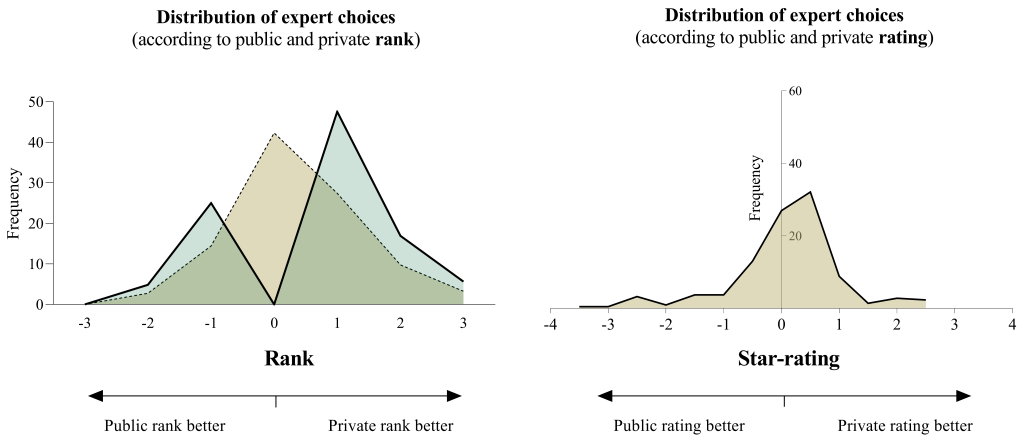


Figure A10: Distribution of selected experts by the spread in public-private rank (left) and rating (right) (neutral frame). For this figure, only the treatment ***Exp+Rating-N*** is considered. **Left**: The dashed line shows the distribution of choices according to ranks including equal ranks. We observed 215 interactions where consumers chose an expert for whom they had both, private experiences, and a public rating. Of those, consumers selected an expert with equal ranks in 42.3% (91 interactions). The solid line (shaded area) only shows the distribution of selected experts when there was a discrepancy between the private and the public ranking. Testing for normality reveals that the distribution is significantly skewed to the left (-0.646, $p < 0.01$). **Right**: we show the distribution of selected experts according to differences between the private and public average ratings (private average rating - public average rating). Hence, positive (negative) numbers indicate that the expert had a better private (public) rating. The distribution, however, is not significantly skewed to the left (-0.356, $p < 0.10$).

A.2 Control condition: Sample comparison

To compare whether study participants at the University of Innsbruck (UIBK) systematically differ from those at the Technical University of Munich (TUM), we ran the experimental condition ***Rating-N*** both in Innsbruck and in Munich, with six independent markets per lab. Regarding background characteristics, the TUM sample has a slightly higher share of men, who are on average older (26.3 versus 21.7). Furthermore, there are fewer students of economics or business (see Table A13).

Regarding behavior in condition ***Rating-N***, we find similar levels of undertreatment, overcharging, and overtreatment, and consequently comparable efficiency levels and consumer surplus (see Table A14). However, it seems that participants in Innsbruck provide feedback less often.

Table A13: Descriptive statistics (subject pool differences: TUM vs UIBK)

	Rating-N (TUM)	*Rating-N (UIBK)*
Male (in %)	54.2	48.8
Age (in years)	26.3	21.7
	(7.1)	(1.0)
Relative School Performance	79.4	74.8
	(17.4)	(22.6)
Number of Physician Visits last year	5.4	4.3
	(8.7)	(4.1)
Exp. with incorrect physician behavior (in %)	25.0	46.5
	(43.8)	(50.5)
Exp. with physician recommendation (in %)	72.9	64.3
	(44.9)	(48.5)
Business/Economics major (in %)	35.4	51.2
	(48.3)	(50.6)
Frequency of practicing Religion (in %)		
Never	54.2	55.8
Rarely	39.6	32.6
Often	6.3	11.6
Extroversion	3.2	3.3
	(1.1)	(0.9)
Agreeableness	3.2	3.4
	(0.8)	(0.9)
Conscientiousness	3.4	3.8
	(0.7)	(0.8)
Neuroticism	2.8	2.8
	(1.0)	(1.1)
Openness	3.5	3.6
	(1.0)	(1.1)
Amount donated to charity in a DG (in euros)	3.7	4.7
	(4.0)	(3.5)
Risk Aversion	12.1	12.7
	(3.1)	(3.3)
Trustworthiness	0.4	0.4
	(0.2)	(0.2)
Liar (in %)	77.1	79.1
	(42.5)	(41.2)
Experimental payoff (physicians)	38.3	42.7
	(21.0)	(24.3)
Experimental payoff (patients)	54.7	46.0
	(20.3)	(13.1)

Note: Means (standard deviations). Background variables were measured in additional experiments and a post-experimental questionnaire (see Appendix F for a detailed description of all these measures). Amount donated to a charity in a dictator game: donation of up to 12 euros as a measure of altruism. Risk aversion: number of safe choices in a choice list with 20 binary decision problems between a risky prospect and a safe option. Trustworthiness: share sent back to the first-mover in a trust game. Liar: dummy variable = 1 if someone reports 4 or more correct dice rolls out of 12 in a lying task. BIG 5 personality traits (extraversion, agreeableness, conscientiousness, neuroticism, and openness) measured with a 10-item BIG 5 questionnaire. Self-reported relative school performance as a proxy for IQ. The experimental payoff is the sum of payoffs in ECUs generated by participants over the 16 periods (not the payout they received at the session's end).

Table A14: Overview of results (means - subject pool effects: TUM vs UIBK)

	TUM	UIBK	*p-values of MWU*[1]
	Rating-N (TUM)	***Rating-N (UIBK)***	***Rating-N (TUM)* vs *Rating-N (UIBK)***
	(1)	(2)	(3)
Expert behavior			
Undertreatment (in %)	8.61	11.73	0.413
Overcharging (in %)	39.86	46.08	0.558
Overtreatment (in %)	0.91	3.15	0.100
Consumer decisions			
Interaction (in %)	100.00	98.44	1.000
Feedback (in %)	96.35	86.48	0.019
Star-rating	3.88	3.54	0.132
Market outcomes			
Efficiency (in %)	96.90	93.63	0.167
Consumer surplus (in ECUs)	3.420	2.79	0.331
Observations	48	48	

Note: We analyze six independent markets in every experimental condition. In each market, four consumers and four experts interact. Please refer to Table 3 for a description of the outcome variables.
[1] Mann-Whitney U-tests for pairwise differences between conditions with matching groups of 8 subjects as one independent observation. p-values are adjusted for the small sample size, using Fisher's exact test.

Table A15: Average treatment effects (subject pool effects: TUM vs UIBK)

	Undertreatment	**Overcharging**	**Efficiency**	**Consumer Surplus**
	(1)	(2)	(3)	(4)
Levels in ***Rating-N (TUM)***	0.132	0.476	0.969	3.419
	(0.025)	(0.076)	(0.010)	(0.366)
	Marginal treatment effects			
Rating-N (UIBK)	0.059	0.006	-0.033*	-0.628
	(0.039)	(0.080)	(0.019)	(0.383)
Period	+***	+***	-*	-**
Observations	371	391	768	768
Number of Groups	12	12	12	12

Note: The table presents results from multilevel models with random effects at the market and individual levels (undertreatment & overcharging: columns 1 & 2) or at the market level for market efficiency (column 5) and consumer surplus (column 6) from conditions with a neutral market frame. See Table 3 for a description of the outcome variables. We report effects as marginal effects, calculated as differences in the expected probabilities between the experimental condition in question and the baseline condition All regressions include time trends. Robust standard errors in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix B Control conditions (Healthcare frame)

In addition to the conditions discussed in the paper, we ran four further conditions, with the healthcare frame, to investigate the role of competition (consumers can choose among 4 experts), personal experience in the absence of competition, and private ratings on market outcomes.

To disentangle the effect of personal experience and competition, we ran two conditions in which consumers were randomly matched with an expert in each round and thus, there was no competition between experts. In one condition, experts are identifiable and consumers can thus attribute personal information to a given expert, (***ExpNoComp-H***) and in the other condition experts are not identifiable (***NoComp-H***). Comparing condition ***Experience-H*** (***Baseline-H***) with ***ExpNoComp-H*** (***NoComp-H***) shows the effect of adding competition in a market with (without) personal experience information, whereas the comparison of ***ExpNoComp-H*** with ***No-Comp-H*** shows the effect of adding personal experience information into a setting without competition between experts.

To disentangle the effect of providing a private rating to experts from the reputational effect of a public rating system, we ran two additional rating conditions in which consumers can rate the interaction with the expert without showing the rating to other market participants (***Rating-Priv-H*** and ***Exp+Rating-Priv-H***). Comparing condition ***Baseline-H*** with ***Rating-Priv-H***, respectively ***Experience-H*** with ***Exp+Rating-Priv-H*** shows the effect of providing feedback (cheap talk) to the expert, whereas the comparison between ***Rating-Priv-H*** with ***Rating-H***, respectively ***Exp+Rating-Priv-H*** and ***Exp+Rating-H*** shows the effect of reputational incentives of the ratings.

Table B1 reports the aggregate results of our main outcome variables for all experimental conditions averaged over markets and periods. The following discussion is based on the results from the regression analysis , which are largely in line with the non-parametric tests.

Competition: In markets without personal experience information, we find that competition (***NoComp-H*** vs. ***Baseline-H***) does not alter market outcomes, except for an unexpected and weakly significant increase in undertreatment. However, the introduction of competition does not result in significantly different levels of market efficiency. If instead, personal experience information is available, allowing consumers to choose among experts significantly improves market outcomes. Undertreatment- and overcharging rates are significantly lower in ***Experience-H*** compared with ***ExpNoComp-H***, leading to higher overall market efficiency. This finding is in line with Huck et al. (2012) who find that only if some form of reputation-building is coupled with competition, market outcomes are enhanced. Similar findings were reported by Brosig-Koch et al. (2017a) and Han et al. (2017), who show that competition among healthcare providers results in higher patient well-being.

Private Feedback: Comparing ***Baseline-H*** with ***Rating-Priv-H***, respectively ***Experience-H*** with ***Exp+Rating-Priv*** allows analyzing the impact of private ratings from consumers to experts. In markets without reputation-building (***Baseline-H***) the mere fact that consumers can send private ratings (***Rating-Priv-H***) to experts significantly decreases undertreatment, whereas there is no effect on overall efficiency. Allowing consumers to give a private rating to experts in markets with personal experience information (***Exp+Rating-Priv-H***) leads to an unexpected but significant increase in overcharging rates while overall market efficiency is not affected.

Reputation effect of ratings: We saw that private ratings seems not to improve market outcomes by and large, except for a reduction in undertreatment in markets with first-time interactions. We have seen, however, that the possibility to rate experts enhances market outcomes when it enables experts to build up a reputation for quality as in condition ***Rating-H***. Comparing ***Rating-Priv-H*** with ***Rating-H***, allows us to analyze the reputational effect of public rating mechanisms. We find highly significant decreases in undertreatment- and overcharging rates, which translate into significantly higher efficiency levels when ratings are made public,

allowing consumers to guide their choice of experts.

Table B1: Overview of all results (means - health frame).

	Markets without personal experience				Markets with personal experience			
	Baseline-H	***Rating-H***	***NoComp-H***	***Rating-Priv-H***	***Experience-H***	***Exp+Rating-H***	***ExpNoComp-H***	***EXP+Rating-Priv-H***
Expert behavior								
Overtreatment (in %)	0.00	4.3	0.60	1.61	6.47	0.56	0.49	0.49
Undertreatment (in %)	64.72	5.81	39.90	42.93	6.81	6.89	24.82	11.76
Overcharging (in %)	92.03	47.94	91.02	86.65	36.89	38.07	88.54	48.51
Consumer decisions								
Interaction (in %)	93.75	98.96	96.88	98.70	99.74	99.48	91.41	100.00
Feedback (in %)	-	93.62	-	84.57	-	88.68	-	77.60
Star-rating	-	3.66	-	3.02	-	4.06	-	3.73
Market outcomes								
Efficiency (in %)	70.70	96.21	80.77	81.06	95.42	96.85	82.53	95.42
Consumer Surplus (in ECUs)	-1.27	3.01	-0.11	-0.10	3.05	3.30	0.55	2.89
Observations	48	48	48	48	48	48	48	48

Note: We analyze six independent markets in every experimental condition. In each market, four consumers and four experts interact.

Appendix C Detailed information about the rating conditions and screenshots

In this section, we describe the information provided to experts and consumers in our rating conditions with a healthcare frame and thus we will refer to the expert as physician and to the consumer as patient. Except for the framing, the information provided in the neutral frame is the same (see Appendix E for the experimental instructions with a neutral frame).

In ***Rating*** physicians see at the end of each period the private rating[43] for patients they treated, and who decided to rate them. Besides, patients (Figure C1) and physicians (Figure C2) observe the public average rating[44] of all physicians over all treated patients when they make their decisions starting in period five. The reason for displaying the public average rating only from period five onwards is to render direct reputation-building impossible in the first couple of rounds, where not many ratings have been submitted so far and identification might be possible via those ratings.

In ***Rating-Priv***, physicians receive at the end of each period a private rating from patients they treated, and who decided to rate them. Neither patients (Figure C3), nor physicians (Figure C4) see any ratings when taking their decisions. Their decision screens look the same as in ***Baseline***.

In ***Exp+Rating*** physicians see at the end of each period the private rating for patients they treated, and who decided to rate them. In line with ***Rating***, patients (Figure C5) and physicians (Figure C6) observe the public average rating of each physician over all treated patients when they make their decisions (from period 2 onwards). Figure C5 shows the decision screen for a patient. However, unlike in condition ***Rating***, physicians also see the private average ratings[45] received from each patient separately, and patients see their own private average ratings from

[43] Private rating from one patient to one physician in any given round on a five-star rating scale. Note that rating a physician is optional in our experiment. Patients may decide not to rate physicians they interacted with.

[44] The public average rating is calculated as the sum of all ratings for a given physician, divided by the number of ratings for this physician.

[45] The private average rating is calculated as the sum of all ratings for a physician by a given patient, divided by the number of ratings the patient has given the physician so far.

previous interactions for each physician separately on top of the average public ratings over all patients. It is important to distinguish between the two average ratings. While the *public average rating* is the average rating for a physician from all patients, the *private average rating* is the average rating for a physician from one patient.

In ***Exp+Rating-Priv***, physicians receive at the end of each period a private rating from patients they treated, and who decided to rate them. Patients see the private average rating for all physicians they rated so far when they decide whether, and which of the physicians to visit (Figure C7). Additionally, physicians observe their private average rating per patient when they decide about treatments and prices (Figure C8).

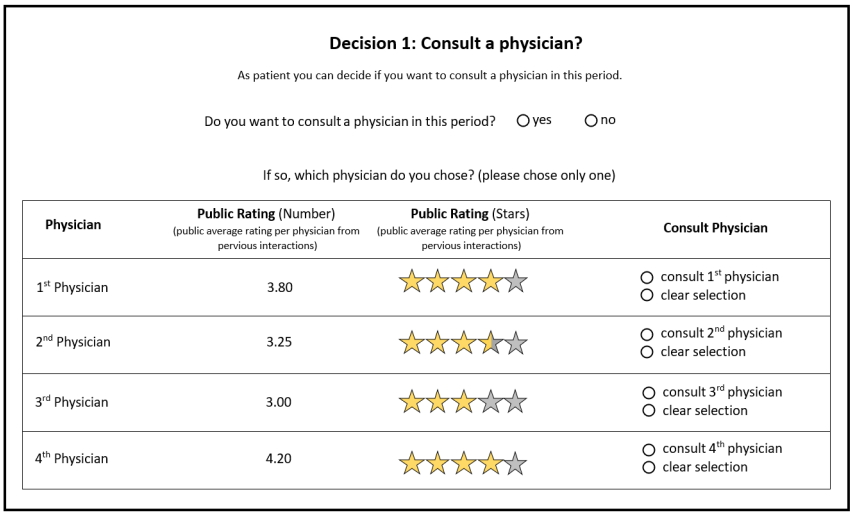


Figure C1: Decision screen of patients in ***Rating***. Patients are asked whether to visit a physician or not. If they do, they may choose one from a list of four. Starting in period five, they see the public average rating once as a number (column 2) and once as a star rating (column 3). These two columns are absent in the first four periods.

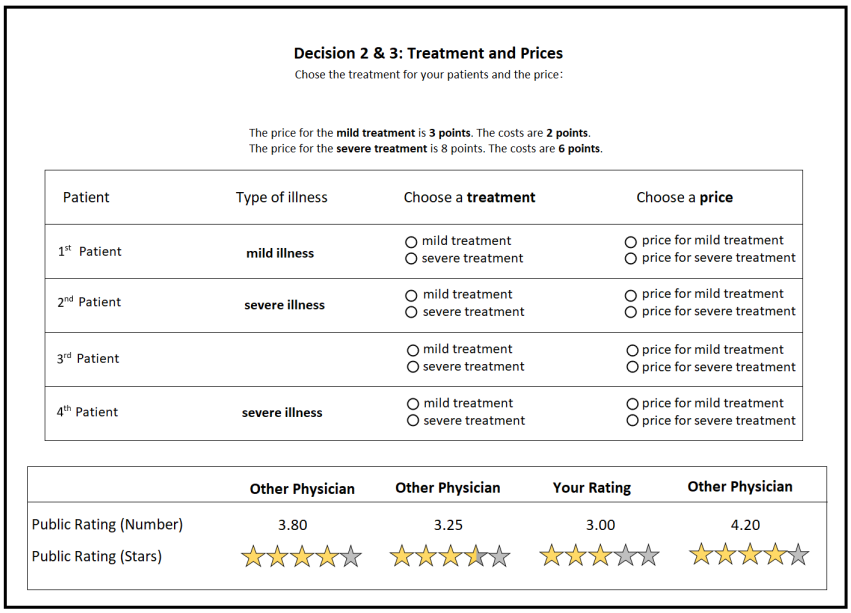
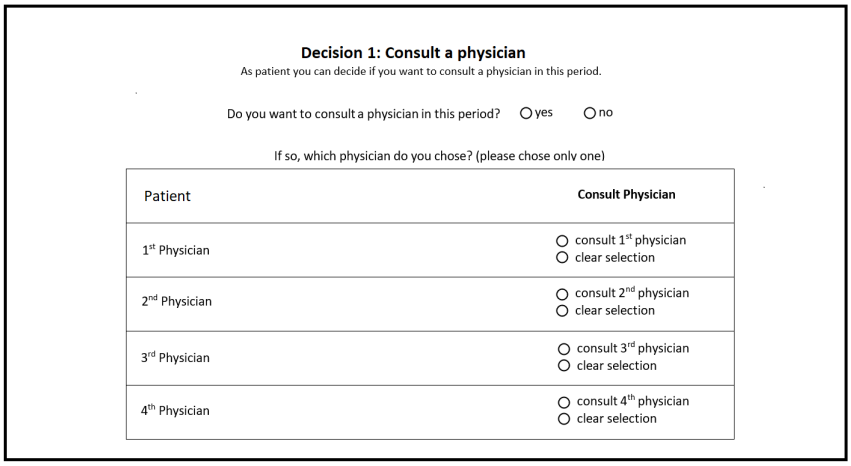


Figure C2: Decision screen of physicians in ***Rating***. Physicians see the type of illness of patients visiting them (column 2). Starting in period five, they see the public rating of themselves and the other physicians. They have to choose a treatment and a price for every visiting patient.



Decision 1: Consult a physician

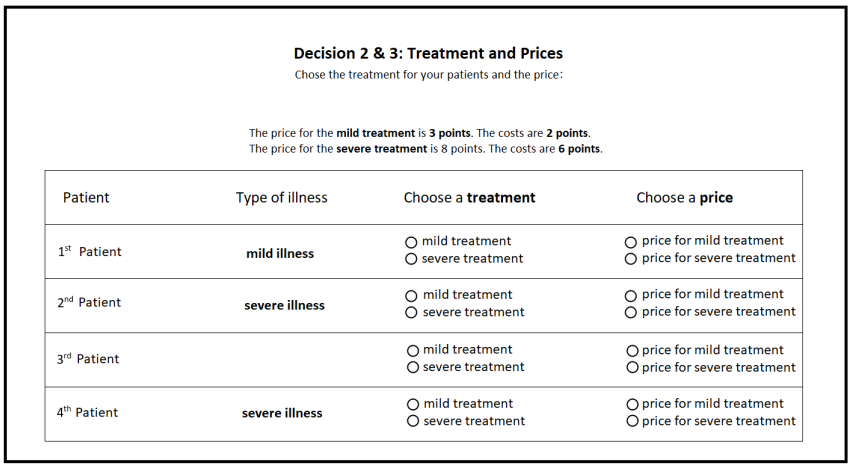
As patient you can decide if you want to consult a physician in this period.

Do you want to consult a physician in this period? ○ yes ○ no

If so, which physician do you chose? (please chose onlv one)

Patient	**Consult Physician**
1st Physician	○ consult 1st physician ○ clear selection
2nd Physician	○ consult 2nd physician ○ clear selection
3rd Physician	○ consult 3rd physician ○ clear selection
4th Physician	○ consult 4th physician ○ clear selection

Figure C3: Decision screen of patients in ***Rating-Priv***. Patients are asked whether to visit a physician or not. If they do, they may choose one from a list of four.



Decision 2 & 3: Treatment and Prices

Chose the treatment for your patients and the price:

The price for the **mild treatment** is **3 points**. The costs are **2 points**.
The price for the **severe treatment** is 8 points. The costs are **6 points**.

Patient	Type of illness	Choose a **treatment**	Choose a **price**
1st Patient	**mild illness**	○ mild treatment ○ severe treatment	○ price for mild treatment ○ price for severe treatment
2nd Patient	**severe illness**	○ mild treatment ○ severe treatment	○ price for mild treatment ○ price for severe treatment
3rd Patient		○ mild treatment ○ severe treatment	○ price for mild treatment ○ price for severe treatment
4th Patient	**severe illness**	○ mild treatment ○ severe treatment	○ price for mild treatment ○ price for severe treatment

Figure C4: Decision screen of physicians in ***Rating-Priv***. Physicians see the type of illness of patients visiting them (column 2). They have to choose a treatment and a price for every visiting patient.

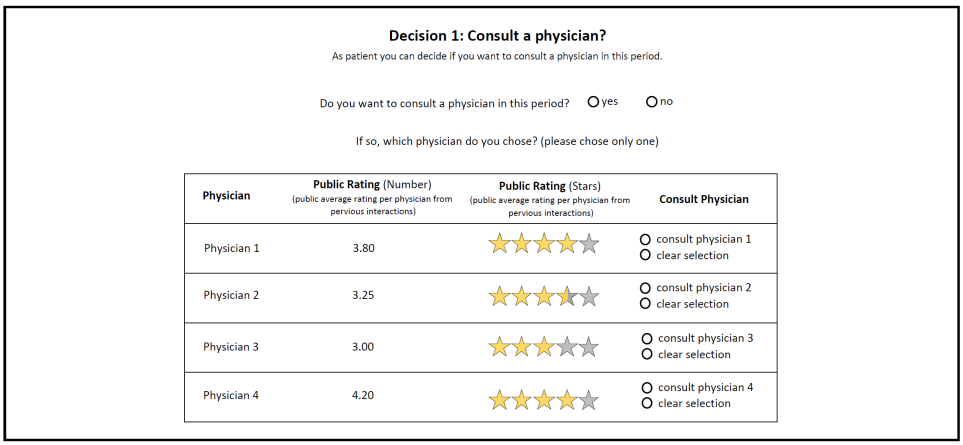


Figure C5: Decision screen of patients in ***Exp+Rating***. Patients are asked whether to visit a physician or not. If they do, they may choose one from a list of four. Starting in period five, they see the public average rating once as a number (column 2) and once as a star rating (column 3). These two columns are absent in the first four periods.

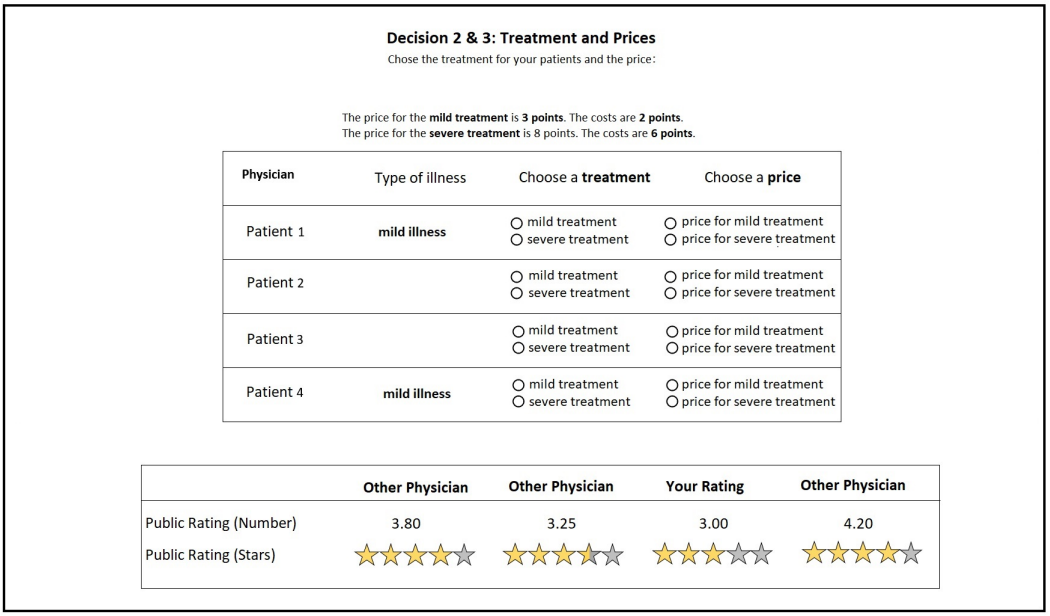


Figure C6: Decision screen of physicians in ***Exp+Rating***. Physicians see the type of illness of patients visiting them (column 4). Starting in period five, they see the public rating of themselves and the other physicians. They have to choose a treatment and a price for every visiting patient.

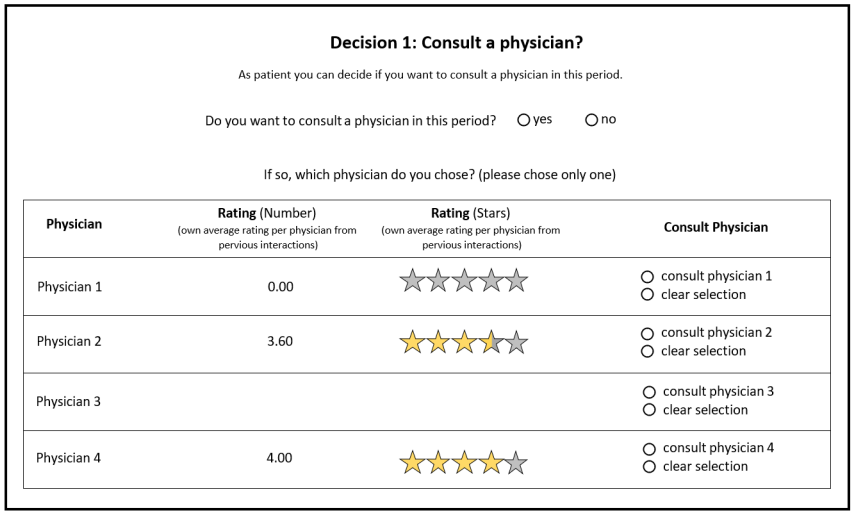


Figure C7: Decision screen of patients in ***Exp+Rating-Priv***. Patients are asked whether to visit a physician or not. If they do, they may choose one from a list of four. Starting in the second period, they see their private average rating for those physicians they already rated as a number (column 2) and a star rating (column 3).

Figure C8: Decision screen of physicians in ***Exp+Rating-Priv***. Physicians see the type of illness of patients visiting them (column 4). Starting in period two, they see the private average rating from patients (columns 2 and 3). They have to choose a treatment and a price for every visiting patient.

Appendix D Predictions

In this section, we construct reputation for quality equilibria for the experimental conditions ***Rating***, ***Experience*** and ***Exp+Rating***.

We assume that consumers and experts are rational, risk-neutral, and maximize their own payoff. All information but the consumers' problem type in a given round is common knowledge. The repeated one-shot equilibrium in which experts provide the minor treatment and charge for the major treatment (price p_H) and consumers always interact is an equilibrium in all experimental conditions. In the following, we construct further symmetric equilibria in which experts do not undertreat/build up a reputation in early periods. The reputation equilibria shown below are not unique, similar ones can be constructed in which the no undertreatment/reputation-building phase is for instance shorter. If necessary, we use masculine pronouns (he) for consumers and feminine pronouns (she) for experts.

Condition *Experience*

Equilibrium without undertreatment in early periods

- Expert's strategy: Provide sufficient treatment (minor treatment for a minor problem, major treatment for a major problem) and charge for the major treatment in periods 1-15. Provide the minor treatment and charge for the major treatment in period 16.
- Consumer's beliefs: Expert provides sufficient treatment and charges for major treatment (price p_H) in periods 1-15. Expert provides minor treatment (q_L) and charges for the major treatment (p_H) in period 16.
- Consumer's strategy: Visit an expert every period. Pick one at random in the first period. In periods 2-15, visit the same expert in the following periods as long as she never undertreated. In periods 1-15, if undertreated, randomly pick one of the experts that never undertreated the consumer before. If there is no such expert, randomly select one. In period 16, choose the expert visited that never undertreated the consumer in any period

1-15. If there is no such expert, choose an expert at random among those never visited before. If there is no expert never visited, randomly select one.

Verification: Consumers' beliefs are consistent with experts' strategies. Next turning to consumers' strategy: In every period it is rational for a consumer to interact as the lowest expected payoff from interaction (in periods 13-16), $0.5 \cdot 2 + 0.5 \cdot (-8) = (-3)$ is higher than the outside option (-4). Furthermore, staying with an expert that never undertreated them is payoff-maximizing. Considering experts, we have to verify that there exists no profitable deviation. In period 16, an expert has no incentive to deviate from her strategy as providing a minor treatment (q_L) and charging for the major treatment (p_H) to every consumer, independently of the health problem of a consumer and the number of consumers, maximizes the expert's payoff. Next, we have to show that sticking to the strategy in periods 1-15 is optimal. In period 15, an expert with a major problem consumer has a continuation payoff of $6 + 0$ from undertreating this consumer, as the consumer will not return given the above strategies, whereas the continuation period from not undertreating is $2+6$ since the consumer will return in period 16. Thus, there is no incentive to deviate to undertreatment in period 15 (for one or more consumers). In earlier periods, the continuation payoff of an expert from undertreatment remains the same, whereas the continuation payoff from sticking to the strategy is even higher as the consumer returns in more periods. Hence, there is no deviation incentive for an expert.

Equilibrium without undertreatment and without full overcharging in early periods

- Expert's strategy: Provide sufficient treatment (minor treatment for minor problem, major treatment for major problem) in periods 1-15, charge for the minor treatment (p_L) in periods 1-11, and charge for the major treatment (p_H) in periods 12-15. Provide the minor treatment and charge for the major treatment in period 16.

- Consumer's beliefs: Expert provides sufficient treatment in periods 1-15. Experts charge for the minor treatment (p_L) in periods 1-11 and charge for the major treatment (p_H) in periods 12-15. Expert provides minor treatment (q_L) and charges for the major treatment (p_H) in period 16.

- Consumer's strategy: Visit an expert every period. Pick one at random in the first period. In periods 2-15, visit the same expert in the following periods as long as she never undertreated in any of the previous periods and always charged p_L in periods 1-11. In periods 1-11, if undertreated or charged p_H, randomly pick one of the experts that never undertreated or charged the patient p_H before. If there is no such expert, randomly select one. In period 16, choose an expert visited that never undertreated the patient in any period 1-15 and never charged p_H in periods 1-11. If there is no such expert, choose an expert at random among those never visited before. If there is no expert never visited, randomly select one.

Verification: Consumers' beliefs are consistent with experts' strategies. Next turning to consumers' strategy: In every period it is rational for a consumer to interact as the lowest expected payoff from interaction (in periods 13-16), $0.5 \cdot 2 + 0.5 \cdot (-8) = (-3)$ is higher than the outside option (-4). Furthermore, staying with an expert that never undertreated them is payoff-maximizing. Considering experts, we have to verify that there exists no profitable deviation. In period 16, an expert has no incentive to deviate from her strategy as providing a minor treatment (q_L) and charging for the major treatment (p_H) maximizes the expert's payoff. Next, we have to show that sticking to the strategy in periods 12-15 is optimal. In period 15, an expert with a major problem consumer has a continuation payoff of $6 + 0$ from undertreating this consumer, as the consumer will not return given the above strategies, whereas the continuation period from not undertreating is $2+6$ since the consumer will return in period 16. Thus, there is no incentive to deviate to undertreatment in period 15 (for one or more consumers). In earlier periods for periods 12-15, the continuation payoff of an expert from undertreatment remains the same, whereas the continuation payoff from sticking to the strategy is even higher as the consumer returns in more periods. Hence, there is no deviation incentive for an expert. In periods 1-11, the expert makes an expected loss per consumer of -1 from not undertreating and always charging p_L, and a loss of -3 with a given major problem consumer. The profit from deviating for a major problem consumer in period 11 is $6 + (16 - 11) \cdot 0 = 6$ which is lower than the continuation payoff from sticking to the strategy with this consumer which amounts to $-3+(15-11)4+6 = 19$. As the expert makes losses in the first periods, deviation

incentives are larger in period 1: The continuation profit from deviating for a major problem patient in period 1 is $6 + (15) \cdot 0 = 6$ which is lower than the continuation payoff from sticking to the strategy with this consumer which amounts to $-3 - 1(11 - 1) + 4(15 - 11) + 6 = 9$. Hence, no expert has an incentive to deviate.

Condition *Rating*

Equilibrium without undertreatment in early periods

- Provision and charging strategy of an expert: Provide sufficient treatment (minor treatment for a minor problem, major for a health problem) and charge for the major treatment in periods 1-13. Provide the minor treatment and charge for the major treatment in periods 14-16.
- Consumer's beliefs: Expert provides sufficient treatment and charges for major treatment (p_H) in periods 1-13. Expert provides minor treatment (q_L) and charges for the major treatment (p_H) in the periods 14-16.
- Consumer's strategy: Visit an expert every period. Randomly pick one expert in periods $1 - 4$. In periods 1-13, give a rating for every interaction following the rule: A rating of 5 stars if the payoff from interaction in the current period is positive, a rating of 0 stars otherwise. Starting in period 5 until period 13, choose randomly among experts with a five-star rating. If, there is no such expert, visit an expert that was never been rated before. If there is no expert with a five-star rating and no expert that was never rated before, pick the highest-rated expert. In periods 14-16, pick the highest-rated expert and do not rate interactions.

Verification: Consumers' beliefs are consistent with the experts' strategy. Next turning to consumers' strategy: In every period it is rational for a consumer to interact as the lowest expected payoff from interaction (in periods 14-16), $0.5 \cdot 2 + 0.5 \cdot (-8) = (-3)$ is higher than the outside option (-4). Furthermore, starting in period 5, it is rational for a consumer to choose the expert with a five-star rating as any rating lower than 5, given the symmetric consumer strategy, signals that this expert undertreated some consumers in earlier periods. For experts, we have to

verify that there exists no profitable deviation from the strategy stated above in any period. In periods 14-16, providing a minor treatment (q_L) and charging for the major treatment (p_H) to a visiting consumer maximizes the expert's payoff. Next, we show that there is also no deviation incentive in any period 5-13: In period 13, an expert with the highest deviation incentives (four major problem consumers) has a continuation payoff of $4 \cdot 6 + 0 = 24$ from undertreating her consumers, as consumers will give a rating of 0 stars and hence there will be no consumers visiting in periods 14-16, as the expert will not have a 5-star rating anymore and given the above-specified strategies of consumers (and other experts). The expected continuation period from not undertreating is $2 \cdot 4 + 6 \cdot 3 = 26$ since the consumers will give a rating of 5 stars and, given the symmetric strategies, the expert will have in expectation one consumer visiting in each of the periods 14-16. Thus, there is no incentive to deviate to undertreatment in period 13. In earlier periods 5-12, the continuation payoff of an expert from undertreatment remains the same, whereas the continuation payoff from sticking to the strategy is even higher as consumers return in more periods. Hence, there is no deviation incentive for an expert in periods 5-13. It remains to show that experts do not deviate in periods 1-4 in which no public ratings are available for expert choice. The incentive to deviate is strongest in period 1, as consumers do not adapt their expert choice in periods 2-4. The maximal deviation profit (undertreating four consumers with a major problem in period 1 and undertreating any consumer thereafter) is $6 \cdot 4 + 3 \cdot 6 = 42$, whereas the continuation payoff from sticking to the above strategy is $2 \cdot 4 + 12 \cdot 4 + 6 \cdot 3 = 74$. Thus, no expert has an incentive to deviate.

Equilibrium without undertreatment and without full overcharging in early periods

- Expert's strategy: Provide sufficient treatment (minor treatment for minor problem, major treatment for major problem) in periods 1-13, charge for the minor treatment (p_L) in periods 1-3, and charge for the major treatment (p_H) in periods 4-13. Provide the minor treatment and charge for the major treatment in periods 14-16.
- Consumer's beliefs: Expert provides sufficient treatment in periods 1-13. Experts charge for the minor treatment (p_L) in periods 1-3 and charge for the major treatment (p_H) in

periods 4-13. Expert provides minor treatment (q_L) and charges for the major treatment (p_H) in periods 14-16.

- Consumer's strategy: Visit an expert every period. Randomly pick one expert in periods $1-4$. In periods 1-3, give a rating for every interaction following the rule: A rating of 5 stars if the payoff from the interaction is 7 (correct treatment, p_L), a rating of 0 stars otherwise. In periods 4-13, give a rating for every interaction following the rule: A rating of 5 stars if the payoff from the interaction is positive, a rating of 0 stars otherwise. Starting in period 5 until period 13, choose randomly among experts with a five-star rating. If, there is no such expert, visit an expert that was never been rated before. If there is no expert with a five-star rating and no expert that was never rated before, pick the highest-rated expert. In periods 14-16, pick the highest-rated expert and do not rate interactions.

Verification: Consumers' beliefs are consistent with experts' strategies. Next turning to consumers' strategy: In every period it is rational for a consumer to interact as the lowest expected payoff from interaction (in periods 14-16), $0.5 \cdot 2 + 0.5 \cdot (-8) = (-3)$ is higher than the outside option (-4). Considering experts, we have to verify that there exists no profitable deviation. In periods 14-16, an expert has no incentive to deviate from her strategy as providing a minor treatment (q_L) and charging for the major treatment (p_H) maximizes the expert's payoff. Next, we have to show that sticking to the strategy in periods 4-13 is optimal. In period 13, an expert with the highest deviation incentives (four major problem consumers) has a continuation payoff of $4 \cdot 6 + 0 = 24$ from undertreating her consumers, as consumers will give a rating of 0 stars and hence there will be no consumers visiting in periods 14-16, as the expert will not have a 5-star rating anymore and given the above-specified strategies of consumers (and other experts). The expected continuation period from not undertreating is $2 \cdot 4 + 6 \cdot 3 = 26$ since the consumers will give a rating of 5 stars and, given the symmetric strategies, the expert will have in expectation one consumer visiting in each of the periods 14-16. Thus, there is no incentive to deviate to undertreatment in period 13. In earlier periods 4-13, the continuation payoff of an expert from undertreatment remains the same, whereas the continuation payoff from sticking to the strategy is even higher as consumers return in more periods. In periods

1-3, the expert makes an expected loss per consumer of -1 from not undertreating and always charging p_L, and a loss of -3 with a given major problem consumer. The incentive to deviate is strongest in period 1, as consumers do not adapt their expert choice in periods 2-4 and experts make losses in early periods. The maximal deviation profit, when facing four major problem consumers, is $6 \cdot 4 + 3 \cdot 6 = 42$, whereas the continuation payoff from sticking to the above strategy is $-3 \cdot 4 - 1 \cdot 2 + 10 \cdot 4 + 6 \cdot 3 = 44$. Thus, no expert has an incentive to deviate.

Condition *Exp+Rating*

Reputation equilibria can be constructed as for *Experience*.

Appendix E Short- and long instructions and control questions for *Rating-H*, *Exp+Rating-H* and *Rating-N*

To save space, we report the instructions for ***Rating-H*** and show the variations for ***Exp+Rating-H*** in brackets and underlined. This is followed by the instructions for ***Rating-N***.

Short-Instructions with helathcare framing (without screenshots)

Problem

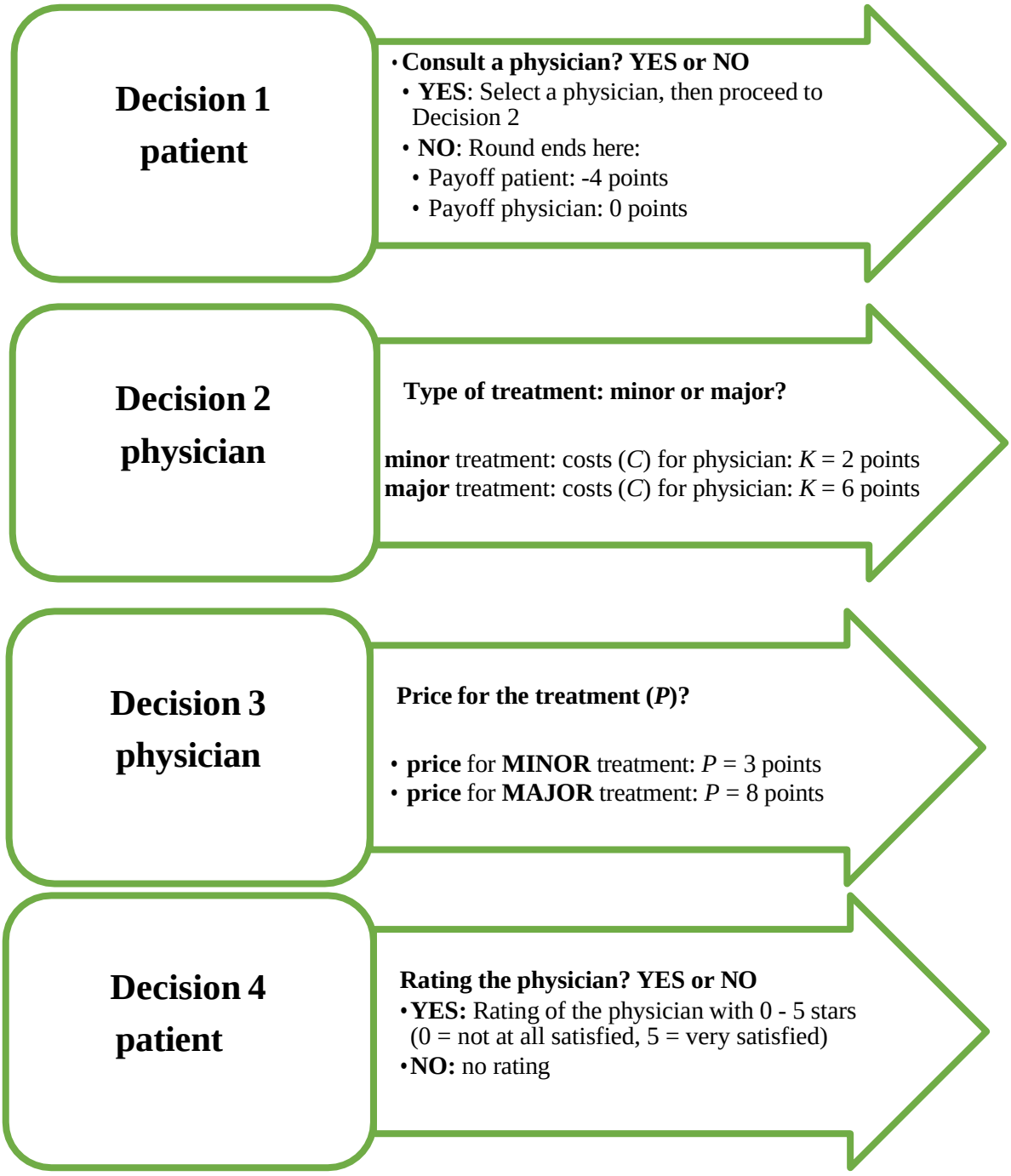
- 16 periods
- 2 roles: **Physician** and **patient**
- Random allocation of the role (remains the same over the entire 16 rounds)
- The patient has an **illness** in every round
- 2 types of illness: **minor** and **major** illness
- **Illness** is randomly **re-determined** in **each round**
- The physician may then freely choose from one of two treatment types: **minor** and **major treatment**
- **NOTE: minor** and **major** treatment **cure minor** illness, **BUT only major** treatment cures **major illness**
  Each round consists of **max. 4 decisions** (see description below)

Information patient

- **The patient** does not know at any time whether he has a minor or major illness in the respective round
- The only information the patient receives is …
  - ... his payoff after decision 2 and 3
  - … if his illness was cured
  - … starting in round 5, the public average rating per physician

Information physician

- **The physician learns** what illness the patient has when the patient decides to go to the physician
- Furthermore, the physician receives information about …
  - ... her payoff per patient according to her decision 3
  - … decision 4 of her patients
  - ... starting in round 5, her own public rating, as well as the public rating of other physicians



Payoff patient

$N - P$

Illness cured: N = 10 points

Illness not cured: N = 0 points

Payoff physician

$P - C$

(Price chosen in decision 3 minus the costs of the treatment chosen in decision 2)

Long-Instructions

Dear participants, welcome to today's experiment!

Please read the instructions for the experiment carefully. All statements in the instructions are true. Your payoff at the end of the experiment depends on how well you have understood those instructions. All data gathered during the experiment will be treated confidentially and evaluated anonymously.

We ask you to remove all items, including other reading materials and writing utensils from the table, and switch off your mobile phone, as well as any other electronic devices. If you have a question, raise your hand and one of the experimenters will come to you to answer your question privately.

All personal designations in this experiment refer equally to men and women.

Thank you very much for your participation in today´s experiment.

Instructions for the experiment

Thank you very much for your participation in the experiment. Please do not speak to other participants until the end of the experiment.

2 roles and 16 rounds

This experiment consists of **16 rounds**, each with the same sequence of decisions. The sequence of decisions is explained in detail below.

There are 2 roles in the experiment: **Physician** and **patient**. At the beginning of the experiment, you will be randomly assigned one of these roles and maintain this role for the entire experiment. On the first screen of the experiment, you can see which role is assigned to you. This role remains the same throughout all periods.

At the beginning of the experiment, you will be randomly assigned to a **group of 7** other players. This **group** remains **the same** for all periods and consists of **4 physicians** and **4 patients**. If you are a patient, the 4 physicians (1st, 2nd, 3rd, and 4th physician) [physician 1, physician 2, physician 3, and physician 4] in your group are your potential interaction partners. If you are a physician, then your potential interaction partners are the 4 patients (1st, 2nd, 3rd, and 4th patients) [patient 1, patient 2, patient 3, and patient 4] in your group. **Note**: The order of the physicians varies randomly from round to round, i.e. the first physician in round one does not necessarily have to be the first physician in round two. The order of the patients varies randomly from round to round as well. [The identification (1, 2, 3, 4) are fixed throughout the experiment, i.e.: A certain patient or physician always has the same identification number (physician 1 is the same person in every round, patient 1 is the same person in every round, etc.)]

All participants receive the same information regarding the rules of the game, including the costs and payoffs for both players.

Overview of the decision situation

Every **patient** is suffering from an **illness** in each period. There are 2 types of illnesses, a **minor** and a **major** illness. Which kind of illness a patient has is determined randomly **each new period**. The patient suffers with a **50% chance** from a **minor illness** and with a **50% chance** from a **major illness**. Imagine a coin toss in each period – if the coin shows "head", then the patient suffers from a minor illness, if it shows "tails", the patient suffers from a major illness. At **no time** is the patient informed whether he has a minor or major illness in a particular round. The physician learns what illness a patient suffers from

only when the patient decides to consult the physician. The physician may then freely choose from one of two treatment types (**minor** or **major** treatment). However, a **major illness** is **only cured** by a **major treatment**. A **minor illness** is **cured** by a **minor or** a **major treatment**.

Overview of the decisions in a round

Each round consists of a maximum of 4 decisions, which are made consecutively. Decision 1 (consult the physician) is made by the patient; decision 2 (treatment) and 3 (price) are made by the physician; decision 4 (rating) is again made by the patient.

The sequence of the decisions of a round and presentation of their consequences

Decision 1

The **patient** decides **whether** he wants to consult **ONE** physician and WHICH of the **4 physicians** (1st, 2nd, 3rd, and 4th physician) he wants to visit (if and with which physician he wants to interact). The order of the physicians is random – at which position a physician appears (as first, second, third, or fourth physician) is determined randomly in each new round.

[The **patient** decides **whether** he wants to consult **ONE** physician and WHICH of the **4 physicians** (physician 1, physician 2, physician 3, and physician 4) he wants to visit (if and with which physician he wants to interact).]

If so, the physician in decision 2 and 3 chooses a treatment and sets a price (see below). However, the patient cannot observe which treatment the physician has chosen.

If not, this round **ends** for the **patient**. **If no patient visits a physician** in a given round, the round ends for her as well.

Decision 2

If the patient decides to consult a physician in decision 1, the **physician learns the nature of the patient's illness** *before* making her decision 2. Then the physician chooses a treatment. At **no time** is **the patient** informed about the treatment chosen by the physician.

The treatment incurs a cost for the **physician**.
The **minor treatment** costs the physician **2 points** (= experimental currency unit) and cures only a minor illness.
The **major treatment** costs the physician **6 points** (= experimental currency unit) and cures both, minor and the major illness.

Physicians can choose treatments **independently** of the type of illness.

Decision 3

The physician **charges** a **price** for the treatment. Two prices are available:

• The price for the **minor treatment** is **3 points**.

• The price for the **major treatment** is **8 points**.

The chosen price **need not** be equal to the price of the treatment chosen in decision 2; it may also be the price of the other treatment.

Decision 4

The patient receives information about his payoff in this round and whether his illness has been cured or not.

Now the patient decides whether he wants to evaluate the interaction with the physician. **If not**, this round ends for him. **If yes**, the patient rates the interaction between 0 (= not satisfied at all) and 5 (= very satisfied) stars.

Afterward, the physician receives information about her payoff and, in case the patient rated her, her rating from this round. The round ends then.

Note: The other physicians and patients also see the ratings: As soon as at least **one** physician was rated by at least **one** patient (i.e. at least one interaction with a physician has been rated), **from the fifth round onwards** all patients see the **public rating** of that physician (i.e. the average value of the ratings from all patients per physician) when asked for their decision 1.

Furthermore, **starting in round** five, physicians see their own **public rating** and the public rating of the other physicians in their group when asked for their decision 2 & 3.

Payoffs

I) No interaction (Patient decides not to consult the physician)

If the **patient** ends the period in decision 1 (decision "**no**" of the patient), then he receives **-4 points** in this period, i.e. he makes a **loss** of 4 points. If **no patient** in a given round **consults a physician**, the round ends for her, and she receives a **payoff** of **zero** points.

Otherwise (decision "**yes**" of the patient) the payoffs are as follows:

II) Interaction (Patient decides to consult the physician)

The **physician** receives the **price** (in points) chosen in decision 3 **minus** the **costs** of the treatment chosen in decision 2 for each of her patients.

For the **patient**, the payoff depends on whether the treatment cured the patient's condition.

a) The treatment has cured the disease. The **patient** receives **10 points minus** the **price** demanded in decision 3.
b) The treatment has not cured the disease. The **patient** must **pay the price** demanded in decision 3.

Two examples to illustrate this:

Example 1:

- The patient decides to consult a physician (Do you want to see a physician in this round = "yes" in decision 1).
- The patient has a major condition.
- The physician chooses a major treatment and charges the price for the major treatment.

Payoff patient: $\underbrace{10}_{\text{benefit treatment}} - \underbrace{8}_{\text{price major treatment}} = 2$

Payoff physician: $\underbrace{8}_{\text{price major treatment}} - \underbrace{6}_{\text{cost major treatment}} = 2$

Example 2:

- The patient decides to consult a physician (Do you want to see a physician in this round = "yes" in decision 1).
- The patient has a minor condition.
- The physician chooses a major treatment and charges the price for the major treatment.

Payoff patient: $\underbrace{10}_{\text{benefit treatment}} - \underbrace{8}_{\text{price major treatment}} = 2$

Payoff physician: $\underbrace{8}_{\text{price major treatment}} - \underbrace{6}_{\text{cost major treatment}} = 2$

The patient and the physician will be informed at the end of each period about their respective payoffs in this period. Besides, the patient learns whether his illness has been cured.

At the beginning of the experiment, you will receive an **initial endowment of 11 points**. You will also receive another **5 points** for **answering the control questions**. From this initial endowment, you can pay for possible losses in individual rounds. Losses can be compensated by winnings from other rounds as well.

At the end of the experiment, four periods will be drawn randomly for payment. For the calculation of payoffs, the initial endowment and the profits or losses over the four payoff-relevant periods are added together. If you have made a total loss at the end of the experiment, you must pay this loss to the experimenter. By participating in the experiment, you agree to this condition. Please note that it is **always** possible to avoid losses in the experiment with certainty. The total number of points will be exchanged for cash at the end of the experiment using the following exchange rate:

1 point = 60 Euro-Cent

(i.e. 5 points = 3 Euro).

You find the experimental receipts on your table. At the end of the experiment, please insert your payoff from the experiment (which you can see on your final screen) on the receipt as well as your first and last name in block letters and sign the receipt.

Control Questions

Here we show the control questions for ***Rating*** and ***Exp+Rating***. Underlined questions represent questions in ***Exp+Rating*** which differ from condition ***Rating***.

It is important to make sure that all participants have fully understood the experiment. Should something has remained unclear, please ask the experimenter. You will receive 5 points (= 3 Euro) for answering the questions correctly. Please answer the following questions:

Question	**Correct Answer**
1. How many decisions does a patient maximally make per period?	2
2. How many decisions does a physician maximally make per period?	2
Assess whether the statements below are true or false.	
3. "The patient learns what illness he suffers from in a particular period."	F
4. "If the physician cures the patient's illness, the total payoff of the patient in this period is exactly 10 points. "	F
5. "Your initial endowment of 11 points is worth 6.60 euros."	T
6. "A physician can identify a patient through the order of line-up over the rounds. That means, for example, that the first patient in the line-up is always the same person."	F
6. "The number of identification (1-4) of patients and physicians are fixed throughout the experiment, i.e. patient (physician) 1 is the same participant in every period."	T
7. "A patient can identify a physician throughout the periods by the order in which they are presented to him, i.e. for example, that the first physician in the list is always the same person."	F
There was no similar question to question 7 in Exp+Rating. The control questions proceeded with question 8 (as question 7)	
8. From the fifth period onwards, all physicians and all patients within the group (of 4 patients and 4 physicians) see the average rating of those physicians already rated as they make their decisions.	T
Please calculate the payoffs for the patient and the physician in the following examples	
9. The patient chooses "No" in decision 1.	Patient: -4 Physician: 0
10. The patient chooses "Yes" in decision 1 and chooses a physician. The patient suffers from a minor illness. The physician chooses a minor treatment and charges the price for a minor treatment.	Patient: 7 Physician: 1
11. The patient chooses "Yes" in decision 1 and chooses a physician. The patient suffers from a minor illness. The physician chooses a minor treatment and charges the price for a major treatment	Patient: 2 Physician: 6
12. The patient chooses "Yes" in decision 1 and chooses a physician. The patient suffers from a major illness. The physician chooses a minor treatment and charges the price for a major treatment.	Patient: -8 Physician: 6

Short-Instructions with neutral frame (without screenshots)

Problem

- 16 periods
- 2 roles: **Player A** and **Player B**
- Random allocation of the role (remains the same over the entire 16 rounds)
- **Player A** has a **problem** in every round
- 2 types of problems: **minor** and **severe** problem
- **Problem** of player A is randomly **re-determined** in **each round**
- Player B offers two different actions: **Action L** and **Action S**
- **NOTE: Action L and action S solve problem L, but only action S solves problem S**
- Each round consists of **max. 4 decisions** (see description below)

Information Player A

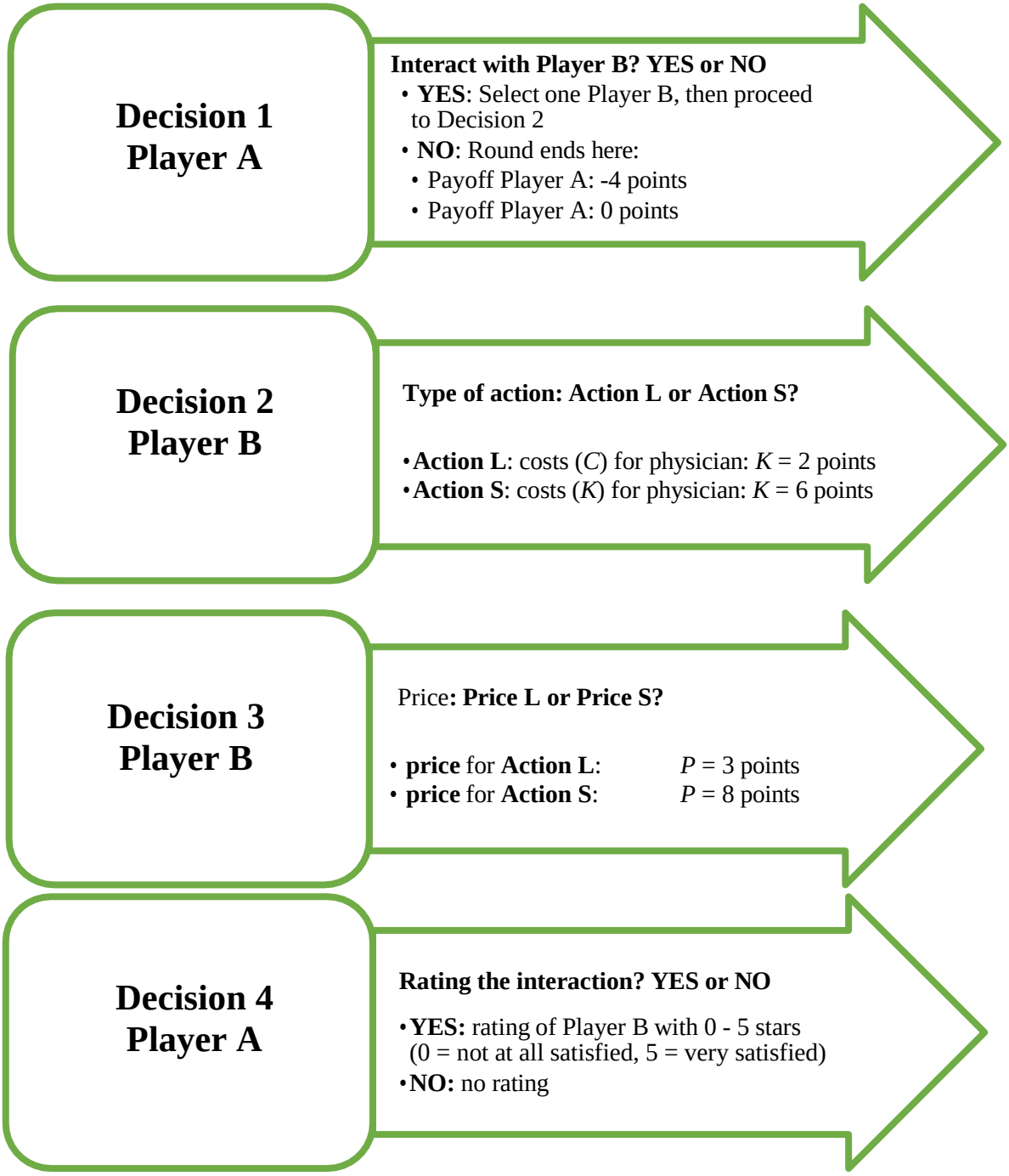
Player A does not know at any time whether he has a minor or major problem in the respective round

The only information Player A receives is …

- ... his payoff after decision 2 and 3
- … if his problem was solved
- … starting in round 5, the public average rating per Player B

Information Player B

- **Player B learns** which problem Player A has (when Player A decides to interact with Player B)
- Furthermore, Player B receives information about …
  - ... her payoff per Player A according to her decision 2 & 3
  - … decision 4 of all the Player A who interact with her
  - ... starting in round 5, her own public rating, as well as the public rating of other Player B



Payoff Player A

N – Price

Problem **solved: *N* = 10 points**

Problem **not solved: *N* = 0 points**

Payoff Player B

Price – Cost

(Price chosen in decision 3 minus the costs of the action chosen in decision 2)

Long-Instructions

Dear participants, welcome to today's experiment!

Please read the instructions for the experiment carefully. All statements in the instructions are true. Your payoff at the end of the experiment depends on how well you have understood those instructions. All data gathered during the experiment will be treated confidentially and evaluated anonymously.

We ask you to remove all items, including other reading materials and writing utensils from the table, and switch off your mobile phone, as well as any other electronic devices. If you have a question, raise your hand and one of the experimenters will come to you to answer your question privately.

All personal designations in this experiment refer equally to men and women.

Thank you very much for your participation in today´s experiment.

Instructions for the experiment

Thank you very much for your participation in the experiment. Please do not speak to other participants until the end of the experiment.

2 roles and 16 rounds

This experiment consists of **16 rounds**, each with the same sequence of decisions. The sequence of decisions is explained in detail below.

There are 2 roles in the experiment: **Player A** and **Player B**. At the beginning of the experiment, you will be randomly assigned one of these roles and maintain this role for the entire experiment. On the first screen of the experiment, you can see which role is assigned to you. This role remains the same throughout all periods.

At the beginning of the experiment, you will be randomly assigned to a **group of 7** other players. This **group** remains **the same** for all periods and consists of **4 Player A** and **4 Player B**. If you are a Player A, the 4 Player B (1^{st}, 2^{nd}, 3^{rd}, and 4^{th} Player B) in your group are your potential interaction partners. If you are a Player B, then your potential interaction partners are the 4 Player A (1^{st}, 2^{nd}, 3^{rd}, and 4^{th} Player B) in your group. **Note**: The order of the players varies randomly from round to round, i.e. the first Player B in round one does not necessarily have to be the first Player B in round two. The order of Players A varies randomly from round to round as well.

All participants receive the same information regarding the rules of the game, including the costs and payoffs for both players.

Overview of the decision situation

Every **Player A** has a **problem** in each period. There are 2 types of problems, a **minor** and a **major** problem. Which kind of problem a Player A has is determined randomly **each new period**. Player A has with a **50% chance** a **minor problem** and with a **50% chance** a **major problem**. Imagine a coin toss in each period – if the coin shows "head", then Player A has a minor problem, if it shows "tails", player A has a major problem. At **no time** is Player A informed whether he has a minor or a major problem in a particular round. Player B learns which problem Player A has when Player A decides to interact with Player B. Player B may then freely choose from one of two actions (**Action L** or **Action S**). However, a **major problem** is **only solved** by **Action S**. A **minor problem** is **solved** by **Action L** and **Action S**.

Overview of the decisions in a round

Each round consists of a maximum of 4 decisions, which are made consecutively. Decision 1 (interact with Player B) is made by Player A; decision 2 (action) and 3 (price) are made by Player B; decision 4 (rating) is again made by Player A.

The sequence of the decisions of a round and presentation of their consequences

Decision 1

Player A decides **whether** he wants to interact with **ONE** player B and WHICH of the **4 Player B** (1st, 2nd, 3rd, and 4th Player B) he wants to interact with. The order of the Players B is random – at which position a particular Player B appears (as first, second, third, or fourth Player B) is determined randomly in each new round.

If so, Player B chooses an action and sets a price in decision 2 and 3 (see below). Player A cannot observe which action Player B has chosen.

If not, this round **ends** for **Player A**. **If no Player A interacts with a Player B** in a given round, the round ends for him as well.

Decision 2

If Player A decides to interact with Player B in decision 1, **Player B learns the type of the patient's problem** *before* making his decision 2. Then Player B chooses an action. At **no time** is **Player A** informed about the action chosen by Player B.

The action incurs a cost for Player B.
Action L costs Player B **2 points** (= experimental currency unit) and solves only a minor problem.
Action S costs Player B **6 points** (= experimental currency unit) and solves both, minor and the major problem.

Player B can choose actions **independently** of the type of problem.

Decision 3

Player B **charges** a **price** for the action. Two prices are available:

- The price for **Action L** is **3 points**.
- The price for **Action S** is **8 points**.

The chosen price **need not** be equal to the price of the action chosen in decision 2; it may also be the price of the other action.

Decision 4

Player A receives information about his payoff in this round and whether his problem has been solved or not.

Now Player A decides whether he wants to rate the interaction with Player B. **If not**, this round ends for him. **If yes**, Player A chooses a rating between 0 (= not satisfied at all) and 5 (= very satisfied) stars.

Afterward, Player B receives information about her payoff and, in case Player A rated her, her rating from this round. The round ends then.

Note: other Player A and Player B also see the ratings: As soon as at least **one** Player B was rated by at least **one** Player A (i.e. at least one interaction with Player B has been rated), **from the fifth round onwards** all Player A see the **public rating** of that Player B (i.e. the average value of the ratings from all Player A per Player B) when asked for their decision 1.

Furthermore, **starting in round** five, Players B see their own **public rating** and the public rating of other Player B in their group when asked for their decision 2 & 3.

Payoffs

I) No interaction (Player A decides not to interact with Player B)

If **Player A** ends the period in decision 1 (decision "**no**" of the Player A), then he receives **-4 points** in this period, i.e. he makes a **loss** of 4 points. If **no Player A** in a given round **interacts with Player B**, the round ends for her, and she receives a **payoff** of **zero** points.

Otherwise (decision "**yes**" of the Player A) the payoffs are as follows:

II) Interaction (Player A decides to interact with Player B)

Player B receives the **price** (in points) chosen in decision 3 **minus** the **costs** of the action chosen in decision 2 for each of the Player A interacting with her.

For **Player A**, the payoff depends on whether the action solved his problem.

a) The treatment has solved his problem. **Player A** receives **10 points minus** the **price** demanded in decision 3.
b) The treatment has not solved his problem. **Player A** must **pay the price** demanded in decision 3.

Two examples to illustrate this:

Example 1:

- Player A decides to interact with Player B (Do you interact with Player B in this round = "yes" in decision 1).
- Player A has a major problem.
- Player B chooses Action L and charges the price for Action L.

Payoff Player A: $\underbrace{10}_{\text{benefit}} - \underbrace{8}_{\text{price Action S}} = 2$

Payoff Player B: $\underbrace{8}_{\text{price Action S}} - \underbrace{6}_{\text{cost Action S}} = 2$

Example 2:

- Player A decides to interact with Player B (Do you interact with Player B in this round = "yes" in decision 1).
- Player A has a minor problem.
- Player B chooses Action S and charges the price for Action S.

Payoff Player A: $\underbrace{10}_{\text{benefit}} - \underbrace{8}_{\text{price Action S}} = 2$

Payoff Player B: $\underbrace{8}_{\text{price Action S}} - \underbrace{6}_{\text{cost Action S}} = 2$

Player A and Player B will be informed at the end of each period about their respective payoffs in this period. Besides, Player A learns whether his problem was solved.

At the beginning of the experiment, you will receive an **initial endowment of 11 points**. You will also receive another **5 points** for **answering the control questions**. From this initial endowment, you can pay for possible losses in individual rounds. Losses can be compensated by winnings from other rounds as well.

At the end of the experiment, four periods will be drawn randomly for payment. For the calculation of payoffs, the initial endowment and the profits or losses over the four payoff-relevant periods are added together. If you have made a total loss at the end of the experiment, you must pay this loss to the experimenter. By participating in the experiment, you agree to this condition. Please note that it is **always** possible to avoid losses in the experiment with certainty. The total number of points will be exchanged for cash at the end of the experiment using the following exchange rate:

1 point = 60 Euro-Cent

(i.e. 5 points = 3 Euro).

You find the experimental receipts on your table. At the end of the experiment, please insert your payoff from the experiment (which you can see on your final screen) on the receipt as well as your first and last name in block letters and sign the receipt.

Control Questions

Here we show the control questions for ***Rating-N*** .

It is important to make sure that all participants have fully understood the experiment. Should something has remained unclear, please ask the experimenter. You will receive 5 points (= 3 Euro) for answering the questions correctly. Please answer the following questions:

Question	**Correct Answer**
1. How many decisions does a Player A maximally make per period?	2
2. How many decisions does a Player B maximally make per period?	2
Assess whether the statements below are true or false.	
3. " Player A learns which problem he has from in a particular period.”	F
4. "If Player B solves Player A 's problem, the total payoff of Player A in this period is exactly 10 points. “	F
5. "Your initial endowment of 11 points is worth 6.60 euros."	T
6. " Player A can identify Player B through the order of line-up over the rounds. That means, for example, that the first Player B in the line-up is always the same person."	F
7. " Player B can identify Player A throughout the periods by the order in which they are presented to him, i.e. for example, that the first Player A in the list is always the same person."	F
8. From the fifth period onwards, all Player A and all Player B within the group (of 4 Player A and 4 Player B) see the average rating of those Player B already rated as they make their decisions.	T
Please calculate the payoffs for Player A and Player B in the following examples	
9. Player A chooses "No" in decision 1.	Player A: -4 Player B: 0
10. Player A chooses "Yes" in decision 1 and chooses one Player B. Player A has a minor problem. Player B chooses Action L and charges the price for Action L.	Player A: 7 Player B: 1
11. Player A chooses "Yes" in decision 1 and chooses one Player B. Player A has a minor problem. Player B chooses Action L and charges the price for Action S.	Player A: 2 Player B: 6
12. Player A chooses "Yes" in decision 1 and chooses one Player B. Player A has a major problem. Player B chooses Action L and charges the price for Action S.	Player A: -8 Player B: 6

Appendix F Experimental instructions for additional games and questionnaire

Part 2:

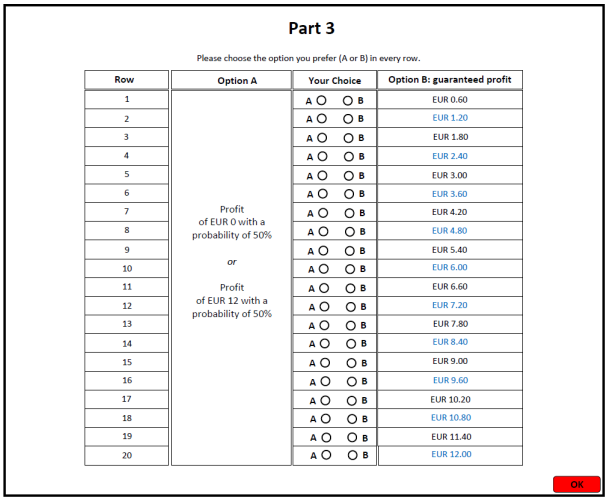
The experiment is not yet over. There are 4 more parts following. At the end of the experiment, one of these parts (part 2, part 3, part 4, or part 5) is randomly selected for payment.

In part 2, you have to make a decision regarding your payoff as well as the payoff of another person. This person is a patient who is supported by the organization "Licht für die Welt". The organization “Licht für die Welt” is known worldwide for preventing and curing preventable blindness. It enables **eye surgery** and **supplies people with eyeglasses and medicines for eye diseases** in South America, Africa, and Asia. You have an endowment of € 12 and you need to decide how you want to divide the money. There are two fields on your screen. One field is marked "amount for me" and the other field is marked "amount for Licht für die Welt". The amounts you enter always have to add up to € 12, in units of € 0.1 (i.e., 10 cents). The transfer will be made online at the end of the experiment. To be able to donate to the organization "Licht für die Welt" correctly, we kindly ask the participant with ID 1 to confirm that the money has been transferred to the organization after the online transfer has been made. As a reminder, this part will only be paid if part 2 is randomly selected for payment at the end of the experiment. This also applies to the donation to "Licht für die Welt".

Part 3:

As a reminder, this part will only be paid if part 3 is randomly selected for payment at the end of the experiment. Part 3 consists of 20 decisions. Below, you are asked to decide for each situation. Each of your choices is a selection between "Option A" and "Option B". "Option A" always offers an uncertain payoff: with a 50% probability, you will receive € 12, and with a 50% probability you receive € 0. "Option B" always offers a safe payoff: with 100% probability you receive an amount that varies from decision to decision (that is, you receive the guaranteed payoff of that row).

The decision situation will be presented to you on the screen as follows:



If part 3 happens to be paid out, one of the 20 decisions (lines) will be randomly selected for payment. Additionally, it will be randomly determined if you won the lottery (you receive € 12) or if you lost the lottery (you receive € 0) (if you have chosen the lottery option). When you have made all decisions, please confirm with "OK".

Part 4:

As a reminder, this part will only be paid if part 4 is randomly selected for payment at the end of the experiment. Part 4 is about guessing the outcome of a die roll in a situation marked by randomness. You play 12 rounds of a dice guessing game. Thereby you should guess the number shown on the dice. The more outcomes you guess correctly, the more money you earn. Each round of the game works as follows:

1. First, guess what number will result from the die roll. If you have a number in your head, press the "Next" button.

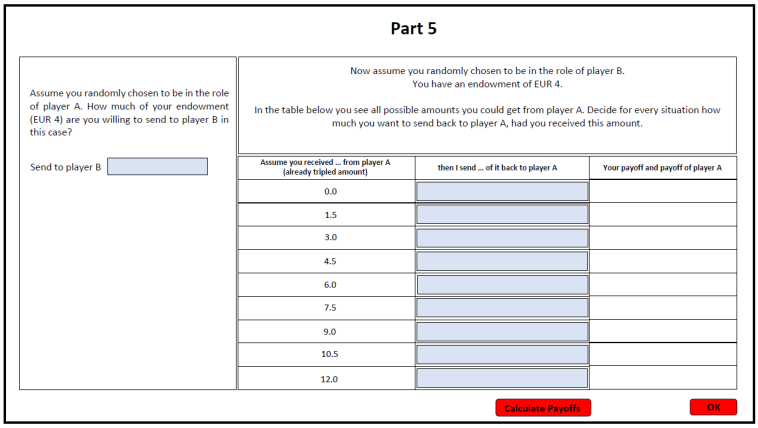
2. Now you see a dice rolled randomly by the computer. Below the dice, you have to enter what number you have guessed.

For each correctly guessed dice you receive 1 €. For each wrongly guessed die roll you receive 20 cents. The profits of all 12 rounds are added up at the end.

Part 5:

As a reminder, this part will only be paid if part 5 is randomly selected for payment at the end of the experiment. Part 5 works as follows: There are two roles, the role of player A and player

B. Both players have an initial endowment of € 4 each. Player A has to decide how much of this endowment (between € 0 and € 4, in 50-cent increments) he wants to send to player B. The total amount sent to player B is tripled. The rest is kept by player A (without tripling). Player B may then decide how much of the tripled amount he wants to send back to player A. You have to decide in the role of player A (see the left side of the decision situation on the screenshot below) as well as in the role of player B (for all possible situations, see the right side of the decision situation on the screenshot below). Only at the end of the game, it will be randomly determined in which role you are in. Besides, you will be assigned to a partner playing the other role. You receive the payoff for your decisions in the role chosen for you at random, in combination with the behavior of your randomly assigned partner.



How well do the following statements describe your personality?

I see myself as someone who …	Strongly disagree	Rather disagree	Neither	Rather agree	Strongly agree
… is reserved	O	O	O	O	O
… is generally trusting	O	O	O	O	O
… tends to be lazy	O	O	O	O	O
… is relaxed, handles stress well	O	O	O	O	O
… has few artistic interests	O	O	O	O	O
… is outgoing, sociable	O	O	O	O	O
… tends to find fault with others	O	O	O	O	O
… does a thorough job	O	O	O	O	O
… gets nervous easily	O	O	O	O	O
… has an active imagination	O	O	O	O	O

Please indicate your gender:

- o Female
- o Male

How old are you?

Which field of study are you in?

Which subject do you study?

(If you are doing several studies, please indicate all and write the study program in parenthesis)

What semester are you in?

What was your average monthly net income over the last year, taking into account all sources of income such as scholarships, student loans, earned income, parental financial support, et cetera? Please round to the nearest ten Euro.

How often do you practice your religion?

- o often
- o rarely

- o never

Please enter here your (average) grade from the Matura/Abitur certificate.

Which grading scale was used in your Matura/Abitur certificate?

- o 1-5 (5 worst rating)
- o 1-6 (6 worst rating)
- o 1-10 (10 best rating)
- o 0 -15 (15 best rating)
- o 0 -100 (100 best rating)
- o other (please specify including explanation)

How do you rate your past average school achievements compared to your former classmates? Answer on a scale from 0 -100 (0 you are the worst student in the class, 100 you are the best student in the class).

How many times have you visited a physician in the last 12 months (including all routine check-ups at the general practitioner, dentist, etc.)?

Have you ever had the impression that a physician is performing more or fewer treatments than necessary or is charging for services that he has not provided?

- o Yes
- o No

Have you ever rated a physician or recommended one?

- o No feedback/recommendation
- o Private feedback to physician
- o Feedback through rating platforms for physicians
- o Recommendation to a friend
- o Other (please specify including explanation)

Have you ever requested a recommendation for a physician?

- o Never requested a recommendation

- o Private recommendation from a friend
- o Looked it up on rating platforms for physicians
- o Other (please specify including explanation)

Were the instructions clear and understandable for you? What could be improved?

Do you have any other comments for us?

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

https://www.uibk.ac.at/eeecon/wopec/

2025-03 **Silvia Angerer, Daniela Glätzle-Rützler, Wanda Mimra, Thomas Rittmannsberger, and Christian Waibel:** The Value of Rating Systems in Credence Goods Markets

2025-02 **Ivo Steimanis, Natalie Struwe, Julian Benda, Esther Blanco:** Reducing strategic uncertainty increases group protection in collective risk social dilemmas

2025-01 **Elias Hasler:** Assessing the Global Impact of EU Carbon Pricing: Economic and Climate Spillovers

2024-13 **Jürgen Huber, Michael Kirchler, Teresa Steinbacher:** Knowledge and Beliefs About Behavioral Biases

2024-12 **Christoph Huber, Felix Holzmeister, Magnus Johannesson, Christian König-Kersting, Anna Dreber, Jürgen Huber, and Michael Kirchler:** Do experimental asset market results replicate? High-powered preregistered replications of 17 claims

2024-11 **Daniela Glätzle-Rützler, Matthias Sutter, and Claudia Zoller:** Coordination games played by children and teenagers: On the influence of age, group size and incentives

2024-10 **Loukas Balafoutas, Esther Blanco, Raphael Epperson:** Targeted Information and Sustainable Consumption: Field Evidence

2024-09 **Armando Holzknecht, Jürgen Huber, Michael Kirchler, Tibor Neugebauer:** Speculating in zero-value assets: The greater fool game experiment

2024-08 **Alexandra Baier, Natalie Struwe:** Accepting the Newcomer: Do Information and Voting Shape Cooperation within Groups?

2024-07 **Max Breitenlechner, Martin Geiger, Mathias Klein:** The Fiscal Channel of Monetary Policy

2024-06 **Silvia Angerer, Hanna Brosch, Daniela Glätzle-Rützler, Philipp Lergetporer, and Thomas Rittmannsberger:** Discrimination in the general population

2024-05 **Rene Schwaiger, Markus Strucks, Stefan Zeisberger:** The Consequences of Narrow Framing for Risk-Taking: A Stress Test of Myopic Loss Aversion

2024-04 **Sebastian Bachler, Armando Holzknecht, Jürgen Huber, Michael Kirchler:** From Individual Choices to the 4-Eyes-Principle: The Big Robber Game revisited among Financial Professionals and Students

2024-03 **IVO STEIMANIS, ESTHER BLANCO, BJÖRN VOLLAN:** Conditional Payments for Democracy to Local Leaders Managing Natural Resources in Rural Namibia

2024-02 **Julian Granna, Stefan Lang, Nikolaus Umlauf:** A Parsimonious Hedonic Distributional Regression Model for Large Data with Heterogeneous Covariate Effects

2024-01 **Philipp Aschersleben, Julian Granna, Thomas Kneib, Stefan Lang, Nikolaus Umlauf, and Winfried Steiner:** Modeling multiplicative interaction effects in Gaussian structured additive regression models

2023-18 **Luke Glowacki, Florian Morath and Hannes Rusch:** High minority power facilitates democratization across ethnic fault lines

2023-17 **Felix Holzmeister, Magnus Johannesson, Robert Böhm, Anna Dreber, Jürgen Huber, Michael Kirchler:** Heterogeneity in effect size estimates: Empirical evidence and practical implications

2023-16 **Changxia Ke, Florian Morath, Sophia Seelos:** Do groups fight more? Experimental evidence on conflict initiation

2023-15 **Loukas Balafoutas, Helena Fornwagner, Rudolf Kerschbamer, Matthias Sutter, and Maryna Tverdostup:** Serving consumers in an uncertain world: A credence goods experiment

2023-14 **Loukas Balafoutas, Helena Fornwagner, Rudolf Kerschbamer, Matthias Sutter, and Maryna Tverdostup:** Diagnostic Uncertainty and Insurance Coverage in Credence Goods Markets

2023-13 **Sarah Lynn Flecke, Rene Schwaiger, Jürgen Huber, Michael Kirchler:** Nature Experiences and Pro-Environmental Behavior: Evidence from a Randomized Controlled Trial

2023-12 **Christian König-Kersting and Stefan T. Trautmann:** Grit, Discounting, & Time Inconsistency

2023-11 **Stefano Piasenti, Marica Valente, Roel van Veldhuizen, Gregor Pfeifer:** Does Unfairness Hurt Women? The Effects of Losing Unfair Competitions

2023-10 **Pascal Kieren, Christian König-Kersting, Robert Schmidt, Stefan Trautmann, Franziska Heinicke:** First-Order and Higher-Order Inflation Expectations: Evidence about Households and Firms

2023-09 **Marica Valente, Timm Gries, Lorenzo Trapani:** Informal employment from migration shocks

2023-08 **Natalie Struwe, Esther Blanco, James M. Walker:** No response to changes in marginal incentives in one-shot public good experiments

2023-07 **Sebastian Bachler, Sarah Lynn Flecke, Jürgen Huber, Michael Kirchler, Rene Schwaiger:** Carbon Pricing, Carbon Dividends and Cooperation: Experimental Evidence

2023-06 **Elisabeth Gsottbauer, Michael Kirchler, and Christian König-Kersting:** Climate Crisis Attitudes among Financial Professionals and Climate Experts

2023-05 **Xiaogeng Xu, Satu Metsälampi, Michael Kirchler, Kaisa Kotakorpi, Peter Hans Matthews, Topi Miettinen:** Which income comparisons matter to people, and how? Evidence from a large field experiment

2023-04 **Tanja Hoertnagl, Rudolf Kerschbamer, Regine Oexl, Rudi Stracke, and Uwe Sunde:** Heterogeneity in Rent-Seeking Contests with Multiple Stages: Theory and Experimental Evidence

2023-03 **Melissa Newham, Marica Valente:** The Cost of Influence: How Gifts to Physicians Shape Prescriptions and Drug Costs

2023-02 **Natalie Struwe, Esther Blanco, James M. Walker:** Determinants of Financial Literacy and Behavioral Bias among Adolescents

2023-01 **Marco Aschenwald, Armando Holzknecht, Michael Kirchler, Michael Razen:** Determinants of Financial Literacy and Behavioral Bias among Adolescents

2022-20 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, and Thomas Rittmannsberger:** Beliefs about social norms and (the polarization of) COVID-19 vaccination readiness

2022-19 **Edward I. Altman, Marco Balzano, Alessandro Giannozzi, Stjepan Srhoj:** Revisiting SME default predictors: The Omega Score

2022-18 **Johannes Diederich, Raphael Epperson, Timo Goeschl:** How to Design the Ask? Funding Units vs. Giving Money

2022-17 **Toman Barsbai, Vojtěch Bartoš, Victoria Licuanan, Andreas Steinmayr, Erwin Tiongson, and Dean Yang:** Picture This: Social Distance and the Mistreatment of Migrant Workers

2022-16 **Andreas Steinmayr, Manuel Rossi:** Vaccine-skeptic physicians and COVID-19 vaccination rates

2022-15 **Stjepan Srhoj, Alex Coad, Janette Walde:** HGX: The Anatomy of High Growth Exporters

2022-14 **Martin Obradovits, Philipp Plaickner** Price-Directed Search, Product Differentiation and Competition

2022-13 **Utz Weitzel, Michael Kirchler** The Banker's Oath And Financial Advice

2022-12 **Julian Granna, Wolfgan Brunauer, Stefan Lang:** Proposing a global model to manage the bias-variance tradeoff in the context of hedonic house price models

2022-11 **Christoph Baumgartner, Stjepan Srhoj and Janette Walde:** Harmonization of product classifications: A consistent time series of economic trade activities

2022-10 **Katharina Momsen, Markus Ohndorf:** Seller Opportunism in Credence Good Markets ? The Role of Market Conditions

2022-09 **Christoph Huber, Michael Kirchler:** Experiments in Finance ? A Survey of Historical Trends

2022-08 **Tri Vi Dang, Xiaoxi Liu, Florian Morath:** Taxation, Information Acquisition, and Trade in Decentralized Markets: Theory and Test

2022-07 **Christoph Huber, Christian König-Kersting:** Experimenting with Financial Professionals

2022-06 **Martin Gächter, Martin Geiger, Elias Hasler:** On the structural determinants of growth-at-risk

2022-05 **Katharina Momsen, Sebastian O. Schneider:** Motivated Reasoning, Information Avoidance, and Default Bias

2022-04 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Thomas Rittmannsberger:** How does the vaccine approval procedure affect COVID-19 vaccination intentions?

2022-03 **Robert Böhm, Cornelia Betsch, Yana Litovsky, Philipp Sprengholz, Noel Brewer, Gretchen Chapman, Julie Leask, George Loewenstein, Martha Scherzer, Cass R. Sunstein, Michael Kirchler:** Crowdsourcing interventions to promote uptake of COVID-19 booster vaccines

2022-02 **Matthias Stefan, Martin Holmén, Felix Holzmeister, Michael Kirchler, Erik Wengström:** You can't always get what you want–An experiment on finance professionals' decisions for others

2022-01 **Toman Barsbai, Andreas Steinmayr, Christoph Winter:** Immigrating into a Recession: Evidence from Family Migrants to the U.S.

2021-32 **Fanny Dellinger:** Housing Support Policies and Refugees' Labor Market Integration in Austria

2021-31 **Albert J. Menkveld, Anna Dreber, Felix Holzmeister, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Sebastian Neusüss, Michael Razen, Utz Weitzel and et al:** Non-Standard Errors

2021-30 **Toman Barsbai, Victoria Licuanan, Andreas Steinmayr, Erwin Tiongson, Dean Yang:** Information and Immigrant Settlement

2021-29 **Natalie Struwe, Esther Blanco, James M. Walker:** Competition Among Public Good Providers for Donor Rewards

2021-28 **Stjepan Srhoj, Melko Dragojević:** Public procurement and supplier job creation: Insights from auctions

2021-27 **Rudolf Kerschbamer, Regine Oexl:** The effect of random shocks on reciprocal behavior in dynamic principal-agent settings

2021-26 **Glenn E. Dutcher, Regine Oexl, Dmitry Ryvkin, Tim Salmon:** Competitive versus cooperative incentives in team production with heterogeneous agents

2021-25 **Anita Gantner, Regine Oexl:** Respecting Entitlements in Legislative Bargaining - A Matter of Preference or Necessity?

2021-24 **Silvia Angerer, E. Glenn Dutcher, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** The formation of risk preferences throughsmall-scale events

2021-23 **Stjepan Srhoj, Dejan Kovač, Jacob N. Shapiro, Randall K. Filer:** The Impact of Delay: Evidence from Formal Out-of-Court Restructuring

2021-22 **Octavio Fernández-Amador, Joseph F. Francois, Doris A. Oberdabernig, Patrick Tomberger:** Energy footprints and the international trade network: A new dataset. Is the European Union doing it better?

2021-21 **Felix Holzmeister, Jürgen Huber, Michael Kirchler, Rene Schwaiger:** Nudging Debtors to Pay Their Debt: Two Randomized Controlled Trials

2021-20 **Daniel Müller, Elisabeth Gsottbauer:** Why Do People Demand Rent Control?

2021-19 **Alexandra Baier, Loukas Balafoutas, Tarek Jaber-Lopez:** Ostracism and Theft in Heterogeneous Groups

2021-18 **Zvonimir Bašić, Parampreet C. Bindra, Daniela Glätzle-Rützler, Angelo Romano, Matthias Sutter, Claudia Zoller:** The roots of cooperation

2021-17 **Silvia Angerer, Jana Bolvashenkova, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Children's patience and school-track choices several years later: Linking experimental and field data

2021-16 **Daniel Gründler, Eric Mayer, Johann Scharler:** Monetary Policy Announcements, Information Schocks, and Exchange Rate Dynamics

2021-15 **Sebastian Bachler, Felix Holzmeister, Michael Razen, Matthias Stefan:** The Impact of Presentation Format and Choice Architecture on Portfolio Allocations: Experimental Evidence

2021-14 **Jeppe Christoffersen, Felix Holzmeister, Thomas Plenborg:** What is Risk to Managers?

2021-13 **Silvia Angerer, Daniela Glätzle-Rützler, Christian Waibel:** Trust in health care credence goods: Experimental evidence on framing andsubject pool effects

2021-12 **Rene Schwaiger, Laura Hueber:** Do MTurkers Exhibit Myopic Loss Aversion?

2021-11 **Felix Holzmeister, Christoph Huber, Stefan Palan:** A Critical Perspective on the Conceptualization of Risk in Behavioral and Experimental Finance

2021-10 **Michael Razen, Alexander Kupfer:** Can increased tax transparency curb corporate tax avoidance?

2021-09 **Changxia Ke, Florian Morath, Anthony Newell, Lionel Page:** Too big to prevail: The paradox of power in coalition formation

2021-08 **Marco Haan, Pim Heijnen, Martin Obradovits:** Competition with List Prices

2021-07 **Martin Dufwenberg, Olof Johansson-Stenman, Michael Kirchler, Florian Lindner, Rene Schwaiger:** Mean Markets or Kind Commerce?

2021-06 **Christoph Huber, Jürgen Huber, and Michael Kirchler:** Volatility Shocks and Investment Behavior

2021-05 **Max Breitenlechner, Georgios Georgiadis, Ben Schumann:** What goes around comes around: How large are spillbacks from US monetary policy?

2021-04 **Utz Weitzel, Michael Kirchler:** The Banker's Oath And Financial Advice

2021-03 **Martin Holmen, Felix Holzmeister, Michael Kirchler, Matthias Stefan, Erik Wengström:** Economic Preferences and Personality Traits Among Finance Professionals and the General Population

2021-02 **Christian König-Kersting:** On the Robustness of Social Norm Elicitation

2021-01 **Laura Hueber, Rene Schwaiger:** Debiasing Through Experience Sampling: The Case of Myopic Loss Aversion.

University of Innsbruck

Working Papers in Economics and Statistics

2025-03

Silvia Angerer, Daniela Glätzle-Rützler, Wanda Mimra, Thomas Rittmannsberger, and Christian Waibel

The Value of Rating Systems in Credence Goods Markets

Abstract
In this paper, we experimentally investigate the effect of public consumer ratings on market outcomes in credence goods markets. Contrary to search or experience goods, consumers cannot evaluate all dimensions of trade for credence goods, which may inhibit the information and reputation-building value of public rating systems. We implement a market in which experts have an informational advantage over consumers with respect to the appropriate service level. The rating system takes the form of a five-star rating system as is common on online rating websites. The value of this rating system is compared in two different expert market settings: First, one in which consumers cannot rely on information from personal experience with the expert, reflecting markets in which consumer-expert interactions are often first-time and infrequent (e.g. specialist visits in healthcare markets). Second, one in which consumers have personal experience with the expert, reflecting markets in which consumer-expert interactions are frequent and repeated (e.g. general practitioner visits in healthcare markets). We find that the public rating system significantly improves market outcomes. Furthermore, a public rating system is a good substitute for personal experience information in terms of market efficiency and consumer surplus. Combined, however, we find no complementarity between public ratings and personal experience information, mainly due to the already high market efficiency in the presence of either one.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)