Are patient-regarding preferences stable?

Evidence from a laboratory experiment with physicians and medical students from different countries

Jian Wang
Department of Health Management and Health Economics, University of Oslo, Dong Furen Institute of Economic and social development ,Wuhan University China

Tor Iversen
Department of Health Management and Health Economics, University of Oslo

Heike Hennig-Schmidt
Department of Economics, University of Bonn and Department of Health Management and Health Economics, University of Oslo

Geir Godager
Department of Health Management and Health Economics, University of Oslo Health Services Research Unit, Akershus University Hospital, Norway

UNIVERSITY OF OSLO
HEALTH ECONOMICS
RESEARCH NETWORK

Working paper 2020: 2

Health Economics Research Network at the University of Oslo
ISSN 1501-9071 (print version.), ISSN 1890-1735 (online), ISBN 978-82-7756-265-0

Are patient-regarding preferences stable? Evidence from a laboratory experiment with physicians and medical students from different countries

Jian Wang[1,2], Tor Iversen[2], Heike Hennig-Schmidt[2,3,4], Geir Godager[2,5,*]

Abstract

We quantify patient-regarding preferences by fitting a bounded rationality model to data from an incentivized laboratory experiment, where Chinese medical doctors, German medical students and Chinese medical students decide under different payment schemes. We find a remarkable stability in patient-regarding preferences when comparing subject pools and we cannot reject the hypothesis of equal patient-regarding preferences in the three groups. The results suggest that a health economic experiment can provide knowledge that reach beyond the student subject pool, and that the preferences of decision-makers in one cultural context can be of relevance in a very different cultural context.

Keywords: Laboratory experiment, Bounded rationality, Payment mechanism, Physician behavior

JEL-Classification: C92, D82, I11, H40, J33

[*] Corresponding author. Email: geir.godager@medisin.uio.no

[1] Dong Fureng Institute of Economic and Social Development, Wuhan University, China

[2] Institute of Health and Society, Department of Health Management and Health Economics, University of Oslo, Norway

[3] BonnEconLab, University of Bonn, Germany

[4] National Research University Higher School of Economics (HSE), Moscow, Russian Federation

[5] Health Services Research Unit, Akershus University Hospital, Norway

Preprint submitted to European Economic Review *February 24, 2020*

1. Introduction

Laboratory experiments provide the opportunity for ceteris-paribus variations, which enable researchers to investigate the causal effects of changes in the variable of interest on behavior (Falk and Fehr, 2003; Falk and Heckman, 2009). Laboratory experiments have potential as a 'test bed' for field experiments and policy reforms, and do not require much time and resources (see Hennig-Schmidt et al. (2011); Cox et al. (2016)). While laboratory experiments have contributed to new knowledge, critics argue that artificial context and specific or irrelevant subject pools might substantially reduce the external validity of results. Recent studies have revealed that findings from many laboratory experiments cannot be replicated, e.g., Camerer et al. (2016), Camerer et al. (2018).

Our study addresses the issues of replicability and validity of experimental results. We "bring the field to the lab" by recruiting medical doctors to our lab experiment by which we study physician decision-making under different payment mechanisms. We use a medically framed setting, where subjects' choices determine both physicians' profit and patients' health benefit. Decisions are incentivized by monetary rewards determined by the payment method in question. Our experiment extends the laboratory experiment of Hennig-Schmidt et al. (2011). In the original between-subject design, subjects were confronted with either capitation (CAP) or fee-for-service (FFS) payment schemes. We extend this study to a within-subject-design, and let each subject decide in both systems.

Doubling the number of decisions, recruiting medical doctors, and conducting the experiment with a substantially larger sample than in previous studies, enables the identification of differences in patient-regarding preferences across subject pools. This paper contributes to the literature by fitting a model of bounded rationality to the incentivized choice data. To the best of our knowledge, this is the first paper to quantify preference parameters in a bounded rationality model using experimental data of medical treatment choices. The large number of choice occasions enables us to quantify the impact of more experienced subjects on the degree of rational decision-making.

We address three research questions. First, we ask whether patient-regarding preferences differ across subject pools. This is an important question addressing the issue of external validity. Recruiting students to participate in experiments is common. If preferences of medical students are different from those of medical doctors, the external validity of student-based results is limited. The few experimental laboratory studies on payment incentives we know of where real doctors are recruited include Brosig-Koch et al. (2016, N=29; 2019, N=104), Fink and Kairies-Schwarz (2019, N=16) and Hafner et al. (2017, N=21). The results are mixed with regard to whether physicians and medical students behave differently. To the best of our knowledge, no previous studies provide parameter estimates of patient-regarding preferences by using a physician sample large enough to provide statistical power in between-subject-pool tests for differences. We estimate preference parameters for physicians from China (N=99), medical students from China (N=178) and medical students from Germany (N=42). We find a remarkable stability in patient-regarding preferences when comparing these subject pools, and we cannot reject the hypothesis of equal patient-regarding preferences in the three groups. Checking the replicability of the results in Hennig-Schmidt et al. (2011) suggests that their findings are robust.

In our second question, we ask how accumulating experience in the lab affects subject behavior. We find that behavior is less random when subjects become more experienced. Within the model context, the interpretation is that experience induces more rational behavior.

Our third research question concerns the validity of results from lab experiments. We ask whether choices by medical doctors in a particular experimental condition can be predicted without using the experimental data on doctors' behavior from this experimental condition. We find that our out-of-sample-predictions of doctors' behavior closely resembles the observed behavior, as the distributions of predicted action probabilities and observed relative frequencies are not significantly different.

The paper proceeds as follows: in Section 2, we relate our study to previous literature. In Section 3, we describe the experimental design, parameters, and procedure. In Section 4 we compare the present experiment with the original study. Section 5 presents an empirical model of bounded rationality, as well as results from maximum likelihood estimation. We discuss our results and conclude in Section 6.

2. Related literature

2.1. Physician payment

The existing literature provides evidence that the design of a payment system for health care providers affects their decisions (see for example Clemens and Gottlieb, 2014; Ellis and McGuire, 1986, 1990; Gosden et al., 2001; Iversen and Luràs, 2000; Iversen, 2004; McGuire, 2000; Ma and Mak, 2019; Scott et al., 2018; Yip et al., 2010; see also Brosig-Koch et al., 2016, 2017; Hennig-Schmidt et al., 2011). A reoccurring result is that FFS —paying for each service provided— promotes activity, and the resulting service volume can be higher than optimal. Likewise, prospective CAP systems encourages the provision of few services, and the resulting service volume can be smaller than optimal (Newhouse, 1996).

Payment systems based on FFS have traditionally been the prevailing payment method for health care providers in many countries. Yet, the rapidly increasing health care expenditures have motivated the discussions of payment reforms, see, for example, Yip and Hsiao (2008); Eggleston (2012). In recent years, policy-makers in many countries (e.g. USA, China, Germany, the Netherlands and Norway) have implemented health care reforms using prospective payment methods including capitation in order to curb the growth in health expenditures.

Most empirical evidence on the effects of payment schemes comes from register- or survey data. Some studies have a quasi-experimental design, as for instance Van Dijk et al. (2013) who make use of the introduction of fee-for-service as a payment component for socially insured consumers in the Netherlands in 2006. The authors find that introducing FFS led to an increase in physician-initiated utilization.

Providing reliable causal inferences about the effects of incentives is challenging with field data, however, due to the potential presence of uncontrolled variation which can include unobserved characteristics of the patient population or self-selection of providers (Gaynor and Gertler, 1995; Sørensen and Grytten, 2003; Devlina and Sarma, 2008).

Only few experimental studies exist investigating the differences between medical students and physicians, and evidence is inconclusive. Among the contributions are Brosig-Koch et al. (2016; 2019). The former study finds that medical students and physicians respond to payment incentives in a qualitatively similar and consistent way. The response differs between subject pools, however, with physicians responding less than students do. In the latter study, the effect on patient-regarding service provision is not significantly different between physicians and medical students.

2.2. Elicitation of preferences

How individuals value available alternatives, and how valuations translates to action are key elements in the analysis of economic choices. There is no consensus on best practice when it comes to representing human behavior by models. However, the assumption that humans maximize an objective has been a fundamental element in the larger part of economic research[6]. One may distinguish between models where observed choices are deterministic, and models where observed choices are the result of a probabilistic process.

Deterministic choices

The work of Paul Samuelson (1938) provides the theoretical foundation for research programs that assume deterministic choices by perfectly rational individuals. A rich literature builds on Samuelson's (1938) *revealed preference principle*, which states that the researcher can infer the preferences of utility maximizing decision-makers based on a sequence of observed choices. The revealed preference (RP) axioms (see, for example Andreoni and Miller (2002) and the references therein) provide necessary and sufficient conditions for a sequence of choices to be consistent with utility maximization. The Weak Axiom of RP, Strong Axiom of RP, and Generalized Axiom of RP, have been subject to rigorous testing by means of field- and lab data. Behaviors violating the RP axioms are frequently found. See, for example, Afriat (1973), Varian (1982,1983), Cox (1997), Mattei (2000), or Février and Visser (2004). By means of Afriat's (1972) *"critical cost efficiency index"* (CCEI) or the related *"violation index"* by Varian (1991), researchers can provide a monetary value of resource waste caused by an individual, or a group of individuals, not behaving according to the theory. Choi et al. (2014) show that inconsistent behavior in laboratory experiments, as measured by CCEI, can predict real world measures such as individual's wealth.

Notable contributions by Tversky and Kahneman (1974; 1979) and Sen (1973; 1977; 1993; 1997) criticized RP theory for being weak on internal consistency and relevance when studying human behavior. Blundell (2005) reviews more recent developments of RP applications, and shows how contemporary methods account for some of the earlier critique. Hands (2013; 2014), argues that the original critique, seems to be less effective against contemporary applications of RP theory than against earlier versions. Empirical methods that rely on RP theory have been applied in health economics, and recent applications include Li et al. (2017; 2018).

Stochastic choices

Some of the critique of empirical revealed preference analysis takes a rather practical perspective. McFadden (1999), for example, considers the perfect rationality assumption of RP theory to be *"unnecessarily strong"*, given the overwhelming contradicting behavioral evidence, and that *"many of the core objectives of economic analysis are attainable with weaker forms of rationality...."* (p. 76).

Assuming choice to be the result of a stochastic process has contributed to substantial achievements in the analysis of economic choices (McFadden, 2001). Recent advances include extensions to strategic decisions

[6] Substantial contributions to the research literature assume bounded rationality under non-maximizing behavior. The work of 1978 Nobel laureate Herbert A. Simon is a notable example. See e.g. Simon (1957, 1979).

(McKelvey and Palfrey, 1995) and choice under uncertainty (Dagsvik, 2008). As described by Dagsvik and Hoff (2011), models applying weaker forms of rationality to allow for inconsistencies and randomness in human behavior are not new to social sciences. Thurstone (1927a; 1927b) is an early contribution. He proposed that even though individuals are able to pick the alternative with the highest utility *at the moment*, utilities vary from moment to moment in a stochastic manner. Thurstone thus describes rational individuals who *act* deterministically and without errors, just like in RP theory. Still, choice becomes a probabilistic process because the utility itself is random. The probabilistic choice models deduced by Luce (1959a) and Tversky (1972) takes a different perspective: The utility of the individual is assumed to be deterministic, while randomness in behavior stems from randomness in agents actions. Luce describes randomness in action caused by individuals' inability to discriminate perfectly between utility levels of available alternatives. While perspectives on the sources of randomness in behavior differ, differences become superficial in practical applications, and McFadden (1981) shows that the two types of probabilistic choice models are equivalent in many cases. The Thurstone-type of models and the Luce and Tversky-type of models are now commonly referred to as random utility models (RUM). The RUM has close links to behavioral models in other fields. According to Glimcher, (2011, p. 72), economic models of random utility can be reduced to psychological models of percept as well as to neurobiological models of biochemical transduction.

We take the Luce and Tversky perspective in this paper. We assume a weak form of rationality, where individuals, who are assumed rational to some *degree*, (behave as if they) maximize a combination of deterministic utility and noise. The bounded rationality model allows for different *degrees* of rationality, and our combination of experimental design and empirical specification enables us to quantify the impact of *experience* in laboratory decision-making on the *degree* of rational decision-making.

We are not the first to study how contextual factors such as experience influence the degree of rationality; see, e.g., Holmes and Boyle (2005), or Olsen et al. (2017). The possible relation between experience in laboratory decision making and rationality in strategic decision making is discussed by McKelvey and Palfrey (1995) who analyze the data by Lieberman (1960), and find strong evidence for a decline in the randomness of behavior when experimental subjects become more experienced in the laboratory.

3. Experiment

3.1. Experimental design

Basic setup and decision situation

The physician in our experiment is assumed to be concerned about her own profit π as well as about the patient benefit B, the latter depending on the quantity of medical services q. The specifics of the experimental design are taken from Hennig-Schmidt et al. (2011). Our experiment differs from theirs, however, in that we apply a within-subject design and let each subject decide in both the CAP and the FFS payment systems, whereas Hennig-Schmidt et al. employ a between-subject setup having different subjects decide in either a CAP or an FFS scheme.

Each participant in our experiment acts in the role of the physician. Their task is to choose a quantity of medical services for a given patient whose health benefit is determined by that choice. Each physician i decides on the quantity of medical services $q \in 0, 1, ..., 10$ for three patient types ($j = 1, 2, 3$) with five abstract illnesses ($k = A, B, C, D, E$). The combination of patient type and illness characterizes a specific patient $1A, 1B, 1C, ..., 3D, 3E$. Patient types differ in the health benefit they gain from the medical services ($B_{1k}(q)$, $B_{2k}(q)$, $B_{3k}(q)$). We use a concave patient benefit function like many theoretical papers do (e.g., Ellis and McGuire, 1986; Ma, 1994; Choné and Ma, 2011). A common characteristic of $B_{jk}(q)$ is a global optimum q^*_{jk} on the quantity interval [0,10] that yields the highest benefit to patients of type j for illnesses k. The level of health benefit patients receive from optimal treatment is nearly the same for all three patient types, only the quantity of medical services differs to get there (see Subsection 3.2 for details). The three types of patients reflect the patients' different states of health (good, intermediate, bad).[7]

The patient health benefit is measured in monetary terms. The physician is sequentially confronted with the same 15 decisions (patients) in both payment systems with either CAP first and FFS second or vice versa.

A physician's choice of medical services simultaneously determines the patient benefit and her own profit ($\pi_{jk}(q)$). The patient is assumed to be passive and fully insured, accepting each level of medical service provided by the physician. In our experiment, no real patients are present. However, physicians' quantity choices have consequences for a real patient outside the lab. The money corresponding to patient benefits aggregated over all decisions was transferred to one real patient's in-hospital account to reduce his out-of-pocket payment for his cancer treatment (see Subsection 3.3 and the instructions in the Supplementary Material, Section C). Thus, subjects have an incentive to care for the patient when making their decisions. We did not inform the participants about the name of the person to whom the money was transferred.

To illustrate the physicians' task, Figure 1a provides the decision screen for patient 1C under FFS whereas Figure 1b shows the decision screen for the same patient under CAP. See also the Chinese screens in the Supplementary Material, Subsection C4. The physician gets information on her remuneration, costs and profit as well as on the patient's benefit for each quantity from 0 to 10. All monetary amounts are in Token, our experimental currency, the exchange rate being 10 Token = 1 RMB for students and 10 Token = 6 RMB for doctors (1 RMB was approximately € 0.12 at the time of the experiment).

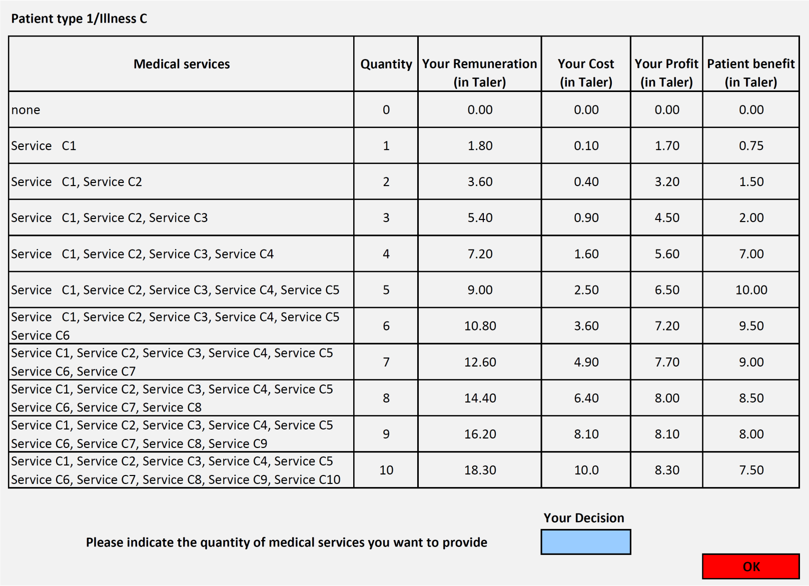
Columns 1 to 6 of the screen, respectively, indicate: (1-2) medical services and the corresponding quantities; (3) physician's remuneration, increasing in the quantity of medical services under FFS (Figure 1a), whereas under CAP the remuneration corresponds to a lump-sum payment per patient (Figure 1b); (4) costs of medical services that are constant across patient types in both parts of the experiment; (5) physician's profit (remuneration minus costs); (6) patient benefit.

3.2. Parameters

To make the studies in China and Germany comparable we kept the specification of the parameters of Hennig-Schmidt et al. (2011) when conducting the experimental sessions in China. Hennig-Schmidt et al. (2011) used the German scale of charges and fees for physician services (Einheitlicher Bewertungsmaßstab) as a guideline for specifying the payment scheme.

[7] Including patients with heterogeneous characteristics in our experiment is motivated by the recent theoretical literature (e.g., Allard et al. 2011), which assumes that patient characteristics affect physicians' behavior. Moreover, empirical findings by Clemens and Gottlieb (2014) indicate that financial incentives have a different impact on physicians' treatment behavior depending on the characteristics of the patients being treated.

Figure 1a: Decision screen for patient 1C under FFS



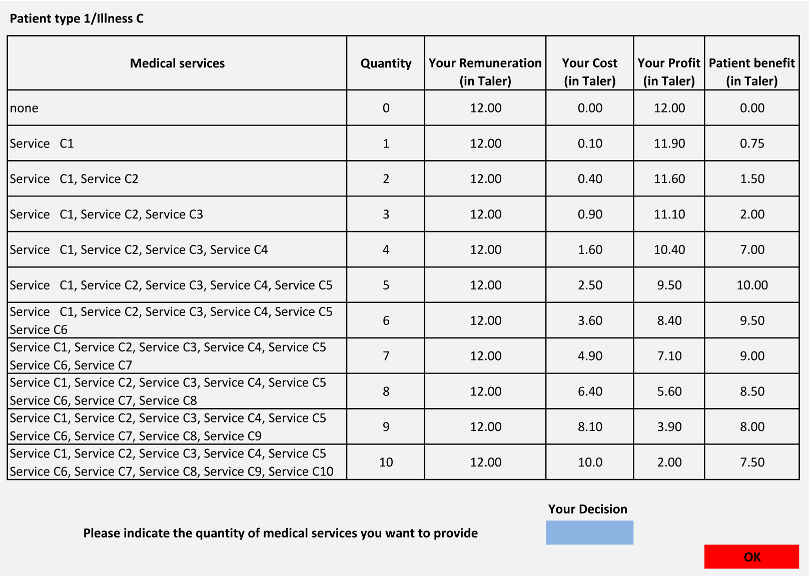
Patient type 1/Illness C

Medical services	Quantity	Your Remuneration (in Taler)	Your Cost (in Taler)	Your Profit (in Taler)	Patient benefit (in Taler)
none	0	0.00	0.00	0.00	0.00
Service C1	1	1.80	0.10	1.70	0.75
Service C1, Service C2	2	3.60	0.40	3.20	1.50
Service C1, Service C2, Service C3	3	5.40	0.90	4.50	2.00
Service C1, Service C2, Service C3, Service C4	4	7.20	1.60	5.60	7.00
Service C1, Service C2, Service C3, Service C4, Service C5	5	9.00	2.50	6.50	10.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6	6	10.80	3.60	7.20	9.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7	7	12.60	4.90	7.70	9.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8	8	14.40	6.40	8.00	8.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9	9	16.20	8.10	8.10	8.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9, Service C10	10	18.30	10.0	8.30	7.50

Your Decision

Please indicate the quantity of medical services you want to provide

Figure 1b: Decision screen for patient 1C under CAP



Patient type 1/Illness C

Medical services	Quantity	Your Remuneration (in Taler)	Your Cost (in Taler)	Your Profit (in Taler)	Patient benefit (in Taler)
none	0	12.00	0.00	12.00	0.00
Service C1	1	12.00	0.10	11.90	0.75
Service C1, Service C2	2	12.00	0.40	11.60	1.50
Service C1, Service C2, Service C3	3	12.00	0.90	11.10	2.00
Service C1, Service C2, Service C3, Service C4	4	12.00	1.60	10.40	7.00
Service C1, Service C2, Service C3, Service C4, Service C5	5	12.00	2.50	9.50	10.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6	6	12.00	3.60	8.40	9.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7	7	12.00	4.90	7.10	9.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8	8	12.00	6.40	5.60	8.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9	9	12.00	8.10	3.90	8.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9, Service C10	10	12.00	10.0	2.00	7.50

Your Decision

Please indicate the quantity of medical services you want to provide

Under FFS, physicians' remuneration increases in q, and remuneration differs with illnesses, $R_{jA}(q), R_{jB}(q), ..., R_{jE}(q)$. Physicians are paid a lump sum of 12 Token per patient under CAP, which was set close to the mean of the maximum profits a subject could achieve under FFS when averaging over patients. For an overview of all payment parameters, see panel I in Table A1 in Appendix A.

The patient benefit $B_{jk}(q)$ varies across patient types. The quantities that maximize patient benefit are $q^*_{1k} = 5$, $q^*_{2k} = 3$ and $q^*_{3k} = 7$ for patient types 1, 2, and 3, respectively with the highest level of health benefit from optimal treatment being nearly the same for all three patient types. Patient benefit $B_{jk}(q)$ is shown in panel IV of Table A1. We refer to quantities smaller than q^*_{jk} as underprovision of medical care, and to provision of quantities larger than q^*_{jk} as overprovision.

Further parameters relevant for physicians' decisions are costs $c_{jk}(q)$ and, particularly, profit $\pi_{jk}(q)$; see panels II and III of Table A.1. Under both payment systems, physicians have to bear costs $c_{jk}(q) = 1/10 \times q^2$. Under CAP, profits are the same for all illnesses, and the profit-maximizing quantity, $\hat{q}$, is 0 for all patients, jk. Under FFS, profits vary across illnesses because remuneration differs while costs are kept constant. The profit-maximizing quantity, $\hat{q}$, is 10 for all patients, jk, except for those with illness A, (i.e., patients 1A, 2A and 3A) as $\hat{q}_{jA} = 5$. For patient 1A, $\hat{q} = q^* = 5$.

The participants are informed on all parameter values before making their treatment decision. For the sake of simplicity, we will in the following number the patients from 1 to 15, keeping in mind that patients 1 to 5 are those of type 1 with an intermediate state of health. Patients 6 to 10 are of type 2 with a good state of health and patients 11 to 15 are of type 3 suffering from a bad state of health.

3.3. Experimental protocol

Applying a within-subject design, each of the 178 Chinese medical students and 99 doctors participating in our experiment was sequentially confronted with the same 15 decisions (patients) in both of the two payment systems FFS and CAP. The subjects were randomly assigned to experimental sessions where either CAP was implemented in Part 1 of the session followed by FFS in Part 2 (condition CF) or in reversed order (condition FC). This design allows us to compare the behavior of the two subject pools over experimental conditions. Each participant was assigned a physician's role and joined the experiment only once, either in CF or in FC. Participants were informed at the beginning that the experiment consisted of two parts, but they did not know what the second part would be.

Our experiment was conducted in September 2012 (medical students) and 2013 (medical doctors) at the Center for Health Economic Experiments and Public Policy at Shandong University in Jinan, China and was programmed with z-Tree (Fischbacher, 2007). All material distributed to the Chinese participants was translated from the original German version by an experienced Chinese translator, being fluent in both Chinese and German.[8] It is important to inform participants in their own language because their behavior may be affected by the language of the instructions ; see e.g. Costa et al. (2014).

Medical students, who voluntarily participated in the experiment, were recruited via notices posted at the campus and by email invitations. Doctors who are working at community health service centers in five districts from north, south, east, west, and the central part of Jinan were recruited through a phone call by the respective District Department of Health informing the doctors that a research experiment from Shandong University needed volunteers. The doctors did not participate during their working hours.

The experimental procedure was exactly the same for medical students and doctors. After having arrived and before the experiment started, participants were randomly allocated to their workstations. The numbered workstations were separated from each other by wooden panels and curtains to guarantee that they made their decisions in anonymity. Then, instructions for Part 1 of the experiment were distributed to participants and read out by a Chinese experimenter. Participants decided under either a CAP or an FFS system. Subjects

[8] The back translation method was applied. For a translation into English, see Supplementary Material, Section C.

were given plenty of time to read the instructions and to ask clarifying questions in private. Questions were answered individually. To check for participants' understanding of the decision task, they had to answer a set of test questions on remuneration, costs, physician profit and patient benefit for a patient they were not confronted with in the actual experiment, see the Supplementary Material, Section C2. Each participant then went through a sequence of 15 choices (patients) on the quantity of medical services to be provided. The order of patients was predetermined and kept constant across conditions. After each decision, each participant in both parts of the experiment was informed about his/her profit and the patient benefit generated by the previous choice. At the end of the first part of the experiment, each participant received information about his/her total profit achieved and the total health benefit generated during all 15 quantity decisions. Finally, the participants answered some open-ended questions.

Next, instructions for the second part of the experiment were distributed and read out by the Chinese experimenter. In Part 2, participants decided under the payment system they had not yet been confronted with. After having completed the second part of the experiment, participants again answered some open-ended questions. The doctors were also asked about socio-demographic variables and professional experience. Next, participants were informed about their individual total profit, the total benefit resulting from their decisions in Parts 1 and 2 of the experiment as well as on their final monetary payoff. Finally, participants were paid in private and dismissed individually.

To ensure that the doctors and medical students trusted the experimenters to actually transfer the money derived from the patient benefit, we used a procedure similar to Eckel and Grossman (1996), Hennig-Schmidt et al. (2011), Godager and Wiesen (2013), Hennig-Schmidt and Wiesen (2014), Godager et al. (2016), Brosig-Koch et al. (2016; 2017; 2019) and Ge et al. (2019). A monitor was randomly selected from the participants in a session. He/she verified the amount of money corresponding to the patient benefits aggregated over all participants' decisions in the respective session. Then, the monitor and an assistant to the experimenters went by taxi to the Shandong University Cancer Hospital in Jinan, and paid the corresponding amount in cash at the hospital-cashier's desk.[9] We took great care to ensure that the monitor did not see the name of the real patient in order to maintain the patient's anonymity. The monitor signed a statement that the appropriate monetary amount was paid into the patient's in-hospital-account. All participants in each session received an email stating the respective amount. Each monitor in the medical student subject pool was paid an additional 50 RMB and each doctor 200 RMB.

We conducted four sessions with medical doctors, and six sessions with medical students. Each experimental session comprised one condition (CF or FC), and lasted for about 90 minutes. Each of the 178 medical students on average earned 28 RMB; 15 RMB (€1.80) in CAP and 13 RMB (€1.56) in FFS plus a show-up fee of 15 RMB (€1.80). Doctors on average earned 160 RMB (86 RMB (€10.32) in CAP and 74 RMB (€8.88) in FFS plus a show-up fee of 120 RMB (€ 14.46).[10] Based on all 8,310 decisions, a total of 19,814 RMB (€2,377.68) was transferred to the real patient's in-hospital-account to be used for reducing his out-of-pocket

[9] We changed the procedure compared to Hennig-Schmidt et al. (2011) who transferred the money to a charity using the money exclusively for cataract surgery. We will discuss the motivations underlying this modification in Section 6.

[10] An acknowledged method in experimental economics of calibrating participants' payoffs—also with regard to cross-cultural comparability—is to adjust stake sizes according to opportunity costs (Herrmann et al., 2008; Gächter and Schulz, 2016). To calibrate the values of experimental tokens, we used the typical hourly wage a participant could earn outside the laboratory. The average payoff for students approximately corresponded to the hourly wage of a student helper at Shandong University of about 30 RMB. For doctors the average hourly wage was about 120 RMB.

payment for cancer treatment; 4,751 RMB (€570.12) for the sessions with medical students and 15,063 RMB (€1,807,56) for the sessions with doctors. Ethical review and approval of the experimental procedure was given by Norwegian Social Science Data Services (reference #44267).

4. Comparing results with the original experiment

We start by describing the subject pools and proceed to testing for differences in aggregate provision behavior between CAP and FFS. Throughout the paper, all statistical tests applied are two-sided. We give a summary of participants' characteristics in Table 1. In our experiment, 277 Chinese subjects participated.[11] Of these, 178 were medical students of whom 56 % were females. The average duration of their medical study was 4.9 semesters. The major of all medical students was Clinical Medicine. The number of participating doctors was 99 with an average age of 40, and 70 % were females. They had on average of 16.23 years of professional experience. The doctors were practicing as general practitioners (75 %), in traditional Chinese medicine (10 %) or in public health (4 %); 11 % of the doctors practiced in all or several of these fields. All doctors were employed at community health centers, where salaries are set according to a fixed scheme. Thus, both the medical students and the doctors have in common that they had little or no practical experience with fee-for-service payment or capitation payment systems.

Table 1: Subject characteristics†

	Chinese Doctors		Chinese students		German students	
Female	70 %	N=99	56 %	N=178	62 %	N=42
Mean Age	40.0	N=89	-		22.3	N=22
Mean semester	-		4.9	N=177	-	
Mean years of practice	16.2	N=88	-		-	

† The German data were provided by Hennig-Schmidt et al. (2011)

Table 2. Aggregate behavior of Chinese doctors and medical students under CAP and FFS. Mean (Std.Dev) of quantity and share of over- and underprovision

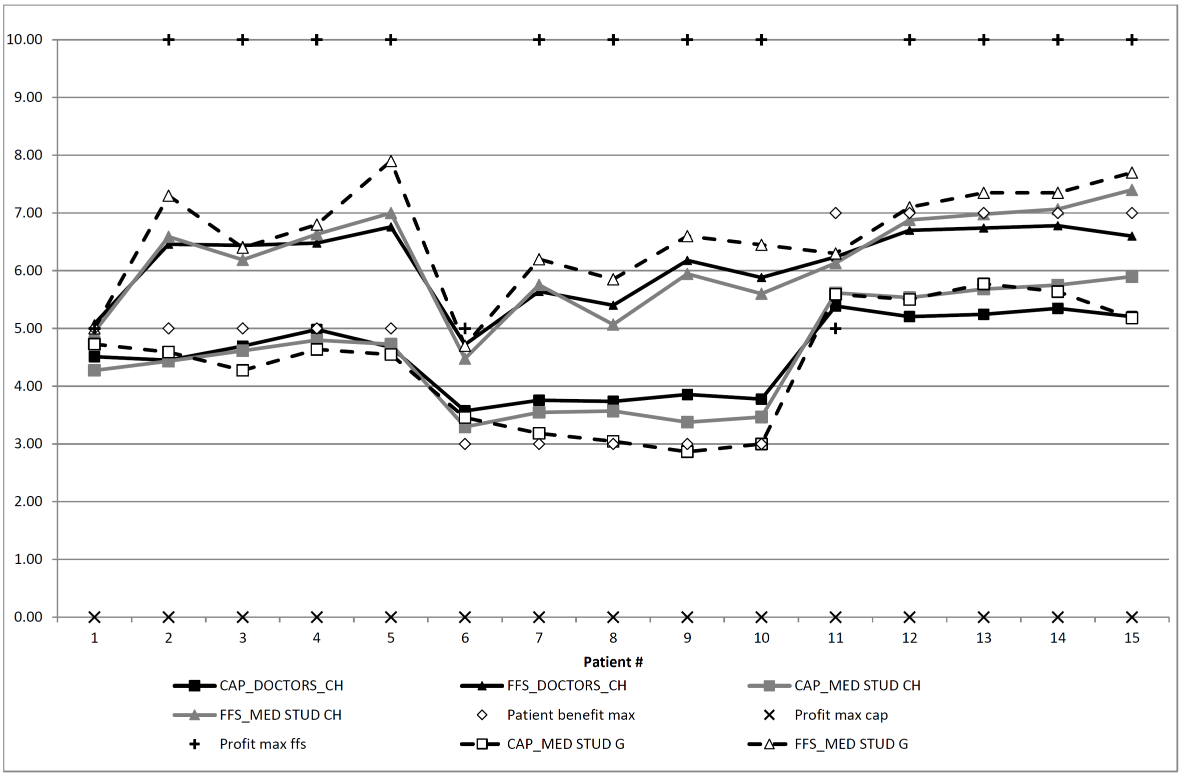
Payment, Variable	Chinese Doctors	Chinese students	Total
	N=1,485	N=2,670	N=4,155
CAP			
Quantity	4.59 (1.78)	4.53 (1.57)	4.55 (1.65)
Overprovision (%)	16	6	10
Underprovision (%)	38	32	34
FFS			
Quantity	6.03 (1.92)	6.16 (1.78)	6.11 (1.83)
Overprovision (%)	49	49	49
Underprovision (%)	16	8	11

Note: This table shows descriptive statistics on quantities of service provision over payment systems and subject pools. #obs is the number of decisions under each payment scheme.

[11] The total number of medical students enrolled in 2012 at the medical campus of Shandong University in Jinan was 400 who were informed via email invitation about the experiment. We thus had a response rate of 44.50%. For doctors, we cannot assess the response rate as we have no information on the number of doctors who had been informed by the District Department of Health on the research experiment at Shandong University. Hennig-Schmidt et al. (2011) achieved a response rate of 37.17%: They invited 113 medical students to participate in their experiment at BonnEconlab, of which 42 participated.

The aggregate provision behavior under CAP and FFS is presented in Table 2. We analyze the data pooled over decisions within the two payment schemes and compare doctors and medical students (N=277 subjects; 4,155 decisions per payment system). We here also pool data from the same payment scheme, regardless of whether the scheme was implemented first or second in the experiment. In line with earlier studies, we find that our participants respond to the incentives given by the payment systems: average quantities in CAP are lower than in FFS (CAP: 4.55, FFS: 6.11; N=277).

Figure 2. Mean quantity provision for each of the 15 Patients under CAP and FFS differentiated according to subject pools – pooled over both parts of the experiment.



Note: This figure shows average quantities of service provision as well as patient benefit and profit maxima for payment systems FFS and CAP for Chinese doctors (N=99), and Chinese medical students (N=178), and German medical students (N=42), pooled over both parts of the experiment.

Our within-subject design enables us to test whether the amount of service provided to a given patient by a given subject differs between the two payment schemes. We conduct 15 tests on the difference between payment schemes, matching the provided service quantity for a given patient in FFS to the corresponding patient scenario in CAP. For each test we may reject the null hypothesis that service quantity does not differ over payment schemes ($p \leq 0.0001$ in each test, Wilcoxon matched-pairs signed-ranks test, WM in the following). Applying a conservative Bonferroni correction for multiple hypothesis testing gives an adjusted threshold for statistical significance of $p = 0.05/15 = 0.0033$ when tests are applied 15 times. Hence, applying Bonferroni corrections would not influence our conclusions.

In addition to the nonparametric analysis, we fit ordinal regression models on q and $B(q)$, and a logistic regression model to analyze how a payment scheme affects the probability of maximizing patient benefit–

which extends the analysis of Hennig-Schmidt et al. (2011). We estimate models with decision-specific fixed effects and individual specific random effects to account for correlation between observations of the same individual, and present the estimation results in Table B0 in the Supplementary Material. The results from these supplementary analyses are consistent with the results provided in Hennig-Schmidt et al. (2011).

As described in Figure (2), overprovision is clearly more prevalent in FFS, for each of the three subject pools. From Figure (2) we also see that under CAP, overprovision for patient type 2 occurs in the experimental sessions conducted in China, while being absent in the German data. Among doctors, 25 to 34 percent of decisions for patient type 2 result in overprovision, and the corresponding figure for students is 8 to 14 percent. Overproviding under CAP is inconsistent with utility maximization, and suggests that assuming perfect rationality is too restrictive. We return to this observation in the following Section 5.

In line with previous studies (Hennig-Schmidt et al., 2011; Keser et al., 2014; Hennig-Schmidt and Wiesen, 2014 and Brosig-Koch et al., 2016, 2017), FFS causes service provision to rise compared to CAP. We conclude that the main findings of Hennig-Schmidt et al. (2011) are confirmed when applying a within-subject configuration of the experiment.

5. Behavioral analysis

5.1. A choice model of bounded rationality

We refer to the vast choice modelling literature that build on the early work of Luce (1959a), Tversky (1972) and McFadden (1974) when deriving our choice model. The conventional way of deriving a choice model as described by Train (2009), is to assume individuals who maximize random utility, and let random utility be the sum of a deterministic utility term and a random term. While the model we derive is a conventional choice model, we want to highlight bounded rationality, and that the degree of rationality depends on the choice situation. We therefore depart somewhat from the conventional formulations, by explicitly assuming boundedly rational individuals who are *unable* to consistently maximize their utility. Differently from McFadden (1974), we assume that behavior is influenced by noise, and that this noise is unrelated to utility. Individuals are assumed to be patient-regarding, deriving utility from both patient benefit, B, and profit, π. Utility is deterministic, and we let $U_n(B_{jt}, \pi_{jt})$ express the utility individual n derives from choosing alternative j in choice occasion t.[12] We introduce bounded rationality, by assuming that individual n chooses alternative j in choice occasion t to maximize an objective $\tilde{F}_{njt}$ given by:

$$\tilde{F}_{njt} = [U_n(B_{jt}, \pi_{jt})]^{\tilde{\lambda}_n} \epsilon_{njt}^{\tilde{\mu}_{nt}} \quad , \tag{1}$$

where the inclusion of the noise terms ϵ_{njt} in the objective implies a departure from utility maximization. The strictly positive parameter $\tilde{\mu}_{nt}$ denotes the weight of the noise term in individual n's objective at occasion t. This parameter measures the behavioral influence of factors that are irrelevant for utility. We consider rationality to always be present to some *degree*, and individuals are, *ceteris paribus*, more likely to maximize their utility when $\tilde{\mu}_{nt}$ is smaller. We note that $\tilde{\mu}_{nt}$ is assumed to vary across choice occasions for a

[12] In our experiment, a *choice occasion* relates to one of the 30 decision screens, 15 in each of the two payment schemes. In order to simplify notation, we will let $t = 1$ ($t = 2$) indicate occasions in the first (second) part of the experiment.

given individual, reflecting that behavior might not be equally affected by noise in all situations. The strictly positive parameter $\tilde{\lambda}_n$ denotes the utility weight in individual n's objective. While assumed constant for individual n, we assume that the utility weights vary between individuals. Individuals are, *ceteris paribus*, more likely to maximize their utility when $\tilde{\lambda}_n$ is larger. Only the *relative* weights of utility and noise in (1) can be identified (Train, 2009), and this relative weight is identified only when appropriate functional form restrictions are introduced for the utility function. The relative noise weight in the objective function (1) is defined by:

$$\sigma_{nt} = \frac{\tilde{\mu}_{nt}}{\tilde{\lambda}_n} \quad . \tag{2}$$

Equation (2) highlights the fundamental identification problem in any behavioral analysis: The fact that only the ratio σ_{nt} can be identified implies that it is not possible to assess whether an individual's randomness in behavior is caused by being particularly responsive to noise that is irrelevant to utility (large $\tilde{\mu}_{nt}$), or caused by a lack of interest in the utility consequences of decisions (small $\tilde{\lambda}_n$). We assume that preferences (α_n) and the subject's interest in the utility consequences of decisions ($\tilde{\lambda}_n$) are fixed during the experiment. Within-subject-differences in σ_{nt}, for example between first and second part of the experimental session, is therefore interpreted as differences in noise influence ($\tilde{\mu}_{nt}$), caused by the variations implemented in the experiment. Importantly however, between-subject-differences in σ_{nt} can be caused by differences in *noise responsiveness*, differences in *utility responsiveness*, or a combination. We return to this fact when discussing the results. In order to identify σ_{nt}, we introduce a functional form restriction by specifying utility to be a Cobb-Douglas function with constant returns to scale:

$$U_n(B_{jt}, \pi_{jt}) = B_{jt}^{\alpha_n} \pi_{jt}^{1-\alpha_n} \quad , \ \alpha_n \in (0,1) \ \forall n \quad , \tag{3}$$

where the fixed parameter α_n is a measure of the relative weight of patient benefit in individual n's utility function. We let the error terms, ϵ_{njt}, in Equation (1) be defined by

$$\epsilon_{njt} = \mathrm{e}^{a_j + \varepsilon_{njt}} \quad , \tag{4}$$

where the ε_{njt} terms are type 1 extreme value distributed, and a_j is a set of alternative specific constants (ASCs).[13] A log transformation of Equation (1) is convenient for discussion and estimation. Inserting (2), (3) and (4) in (1), taking logs and rearranging, our model can be written:

$$F_{njt} = \alpha_n ln(B_{jt}) + (1-\alpha_n) ln(\pi_{jt}) + \sigma_{nt}[a_j + \varepsilon_{njt}] \quad , \tag{5}$$

where $F_{njt} = ln(\tilde{F}_{njt})/\tilde{\lambda}_n$. We henceforth refer to σ_{nt} as the *scale parameter*[14]. From (5) we see that σ_{nt} is inversely related to the *degree of rationality*, as behavior becomes consistent with utility maximization in the limit where σ_{nt} approaches zero. The model in Equation (5) is a so-called scaled multinomial logit model (S-MNL). The S-MNL model of Fiebig et al. (2010) allows for a log-normally distributed inverse scale

[13] ASCs relax the assumption of *independence of irrelevant alternatives*. Following Fiebig et al. (2010), the alternative specific constants are not scaled. The reason is that alternative specific constants are fundamentally different from observable attributes, and it is reasonable to consider ASCs to be part of the error structure.

[14] We follow the terminology and notation in Train (2009). Train (2009, p 40-41) refers to the σ as the *scale parameter*. What Fiebig et al. (2010) refer to as the *scale of the error term* on page 397, right column, corresponds to σ^{-1} in Train's notation, and to the *rationality parameter* λ in McKelvey and Palfrey (1995). We refer to σ^{-1} as the *inverse scale parameter*.

parameter given by $\sigma_{nt}^{-1} = exp(\theta z_{nt} + \tau\eta_n)$, where z_{nt} is a vector of variables which vary over n and t, but are constant within each choice occasion, and $\tau\eta_n$ captures stochastic heterogeneity in scale. As highlighted by Hess and Rose (2012) and Hess and Train (2017), it is not feasible to identify stochastic heterogeneity in both scale (σ_{nt}) and attribute taste (α_n).[15] Our aim is, fortunately, less ambitious. We identify the impact of decision-screen dummies and laboratory experience on σ_{nt}, under the assumption that preferences (α_n) are fixed and independent of the decision task. We assume that scale is determined by observable variables. As noted by Fiebig et al. (2010), choice modellers often lack relevant data for modelling heterogeneity caused by observables, and flexible random coefficient models that account for unobservable heterogeneity are most commonly applied. Our experimental data enables us to identify and quantify how observables describing the decision situation affect the subjects' degree of rationality. Stochastic heterogeneity in scale and preferences are assumed absent.[16] The inverse scale parameter in our empirical specification is given by:

$$\sigma_{nt}^{-1} = exp(\theta z_{nt}) \quad . \tag{6}$$

Included in z_{nt} are a constant term, two dummies equal to 1 for correspondingly medical doctors and German students (meaning that Chinese medical students is the reference category), a dummy equal to 1 in choice occasions where subjects are experienced ($t = 2$), and 17 dummies which indicate the 18 unique choice occasions, 15 in FFS and 3 in CAP[17].

Our model does not not impose a strong rationality assumption. For example, an individual might choose a Pareto-inferior alternative, for example by overproviding services under CAP payment, as reported in Section 4. Also, an individual might choose A rather than B in one occasion, and B rather than A in another, identical occasion. Such behavior would be inconsistent with a strong rationality assumption. Our application of the S-MNL model relies on the assumption that some degree of randomness in behavior is present. Before proceeding to the estimation, we note that the hypothesis that subject behavior is influenced by randomness can be supported by data directly: Under CAP, each subject makes treatment decisions five times for each patient type without any variation in incentives. We find that subjects in all three subject pools frequently make different choices across identical scenarios. For patient 1, 146 (49 %) subjects make the same treatment choice in each of the 5 identical choice occasions, whereas 153 subjects (51 %) vary their treatment choice and are observed with more than one unique action. Correspondingly, 115 (38%) and 186 (62%) subjects vary their treatment choice for patients 2 and 3 (see Table B1 in the Supplementary Material). With this finding in mind, we assume that $\sigma_{nt} > 0$ when we estimate the parameters α_n and θ.

5.2. Estimation and results

In the experimental protocol of Hennig-Schmidt et al. (2011) and in our experiment, the real values of the experimental tokens were set with the aim that hourly payment rates within the experiment are close to subjects' alternative income. In the estimations that follow, we use the experimental token *as is*, without

[15] As noted by Hess and Train (2017), the most flexible model is a mixed logit model where scale is constrained and correlations between preference parameters are allowed.

[16] Our model is thus a restricted version of the S-MNL model discussed by Fiebig et al. (2010), as the τ parameter is fixed to zero.

[17] The payment is a constant in CAP, and only the type of patient can differ between screens. Therefore, the CAP condition includes only three unique screens, one for each patient. For each of the three patients, the subject is confronted with five identical decision screens.

converting it to any real currency. It can be shown that the choice of token exchange rate is irrelevant given the Cobb-Douglas specification and that σ_{nt} is allowed to vary between groups with different token exchange rates.[18]

In the experimental design, some available alternatives have either zero profit or zero patient benefit, which complicates the use of logs. This is solved by replacing ln(0) by 0, and introducing a dummy equal to 1 if either profit or patient benefit is zero. We estimate the parameters of the S-MNL model by means of STATA 15 (Gu et al., 2013), and present the estimation results in Table 3.

Table 3: Results from maximum likelihood estimation
Sample: 178 Chinese students, 99 Chinese doctors, 42 German students. 30 (15) decisions for each Chinese (German) subject. Subjects are more experienced when $t = 2$.

		Chinese student		**Chinese doctor**		**German student**	
α_n		0.51 * CI(0.36 -0.66)		0.42* CI(0.29 - 0.55)		0.40* CI(0.23 - 0.58)	
		t=1	t=2	t=1	t=2	t=1	t=2
σ_{nt}†	FFS$_1$	0.31	0.19	0.61	0.37	0.23	(N.A)
	FFS$_2$	0.37	0.23	0.73	0.45	0.28	
	FFS$_3$	0.35	0.21	0.68	0.42	0.26	
	FFS$_4$	0.32	0.20	0.64	0.39	0.24	
	FFS$_5$	0.41	0.25	0.82	0.50	0.31	
	FFS$_6$	0.14	0.09	0.28	0.17	0.11	
	FFS$_7$	0.46	0.28	0.90	0.55	0.34	
	FFS$_8$	0.29	0.18	0.58	0.35	0.22	
	FFS$_9$	0.39	0.24	0.76	0.47	0.29	
	FFS$_{10}$	0.57	0.35	1.13	0.69	0.43	
	FFS$_{11}$	0.27	0.17	0.54	0.33	0.21	
	FFS$_{12}$	0.36	0.22	0.70	0.43	0.27	
	FFS$_{13}$	0.20	0.12	0.40	0.24	0.15	
	FFS$_{14}$	0.29	0.18	0.58	0.36	0.22	
	FFS$_{15}$	0.20	0.12	0.40	0.24	0.15	
	CAP$_{1\text{-}5}$	0.55	0.34	1.08	0.66	0.41	
	CAP$_{6\text{-}10}$	0.49	0.30	0.96	0.59	0.37	
	CAP$_{11\text{-}15}$	0.23	0.14	0.46	0.28	0.18	

Note: Standard errors are clustered at the level of each individual when computing CI.
* Estimated parameter is significantly different from zero with a p-value < 0.001
† Based on estimated θ parameter, σ_{n1} is significantly different from σ_{n2} with a p-value < 0.001

We compute the σ_{nt}-estimates, which are specific to subject-type and occasion, by inserting the estimated θ-vector in (6). In the following, we let $n = c$ denote Chinese medical student, $n = d$ denotes Chinese medical doctor, and $n = g$ denote German medical student.

[18] Invariance to the unit of measurement is discussed thoroughly by Luce (1959a; 1959b). If included explicitly in (1), token exchange rate would simply be an additive, subject-specific constant in (5). In the event that more valuable tokens cause subject n to become more interested in the utility consequences of decisions, this would be captured by σ_{nt} being specific to n.

The estimated values of α_n ranges from 0.40 (German medical students) to 0.51 (Chinesen doctors). The confidence intervals of α_c, α_d and α_g in Table 3 have substantial overlap.[19] We test the joint hypothesis $\alpha_c = \alpha_d = \alpha_g$, and find that this hypothesis cannot be rejected (p-value 0.28, Wald tests). With reference to our first research question, we do not find any evidence suggesting that patient-regarding preferences differ between subject pools. Preferences of German and Chinese subjects are not significantly different, and we do not reject the hypothesis that preferences are stable in space. We also do not reject the hypothesis that preferences are stable over age, as preferences of medical students and medical doctors are not significantly different.

RESULT 1: We do not find evidence that patient-regarding preferences differ between subject pools.

We observe that the point estimates of the scale parameters of Chinese medical doctors in Table 3 are generally larger than for medical students. While the interpretation is that the behavior of medical doctors is more random than the behavior of students, we cannot assess whether the differences in σ_{nt} are caused by differences in *noise responsiveness* or differences in *utility responsiveness*, since σ_{nt} is a ratio. Hence, it is unknown whether the medical doctors act more randomly because they are particularly responsive to noise, or because they are less interested in the utility consequences of their decisions in the experiment, or a combination of the two. The levels of profits and patient benefits, as well as the mechanism that maps choices to profits and patient benefits, differ between choice occasions. Differences in σ_{nt} between choice occasions for the same individual should be expected. We see in Table 3 that for all three subject pools, the point estimates for the scale parameter is generally larger under CAP than FFS, and the interpretation is that CAP causes more randomness in behavior than FFS.

We find strong evidence that *experience* (t=2) causes reductions in σ_{nt}. Our interpretation is that experience causes more rational behavior. We see that for the Chinese subject pool, with an additional second payment scheme adding 15 choice occasions to the experiment, the influence of noise on decision-making is reduced in occasions where subjects are experienced ($t = 2$) compared to when they are inexperienced ($t = 1$). This implies that subjects are significantly more likely to choose their utility maximizing response in ($t = 2$) compared to ($t = 1$) (McKelvey and Palfrey, 1995). The hypothesis that experience does not affect the degree of rationality can be rejected for both Chinese medical students and medical doctors - the two subject pools who experienced both $t = 1$ and $t = 2$.

RESULT 2: We find evidence that experience increases the degree of rationality in decision-making.

5.3. Examining the validity of results.

We now show that the use of student subjects in lab experiments can contribute to knowledge on how medical doctors *would* behave in a similar situation. Based on the result that preferences of students and medical doctors are not statistically different, we refit a restricted version of our model, by constraining

[19] Point estimates of α_n for German medical students and Chinese doctors are comparable in magnitude to results reported by Li (2018), who analyze data from an experiment with neutral framing.

preferences to be identical across subject types, assuming $\alpha_n = \alpha \ \ \forall n$. We fit this model on two subsets of data:

- When predicting behavior of Chinese medical doctors in CAP, all information on their behavior in CAP is excluded from our estimation.
- When predicting behavior of Chinese medical doctors in FFS, all information on their behavior in FFS is excluded from our estimation.

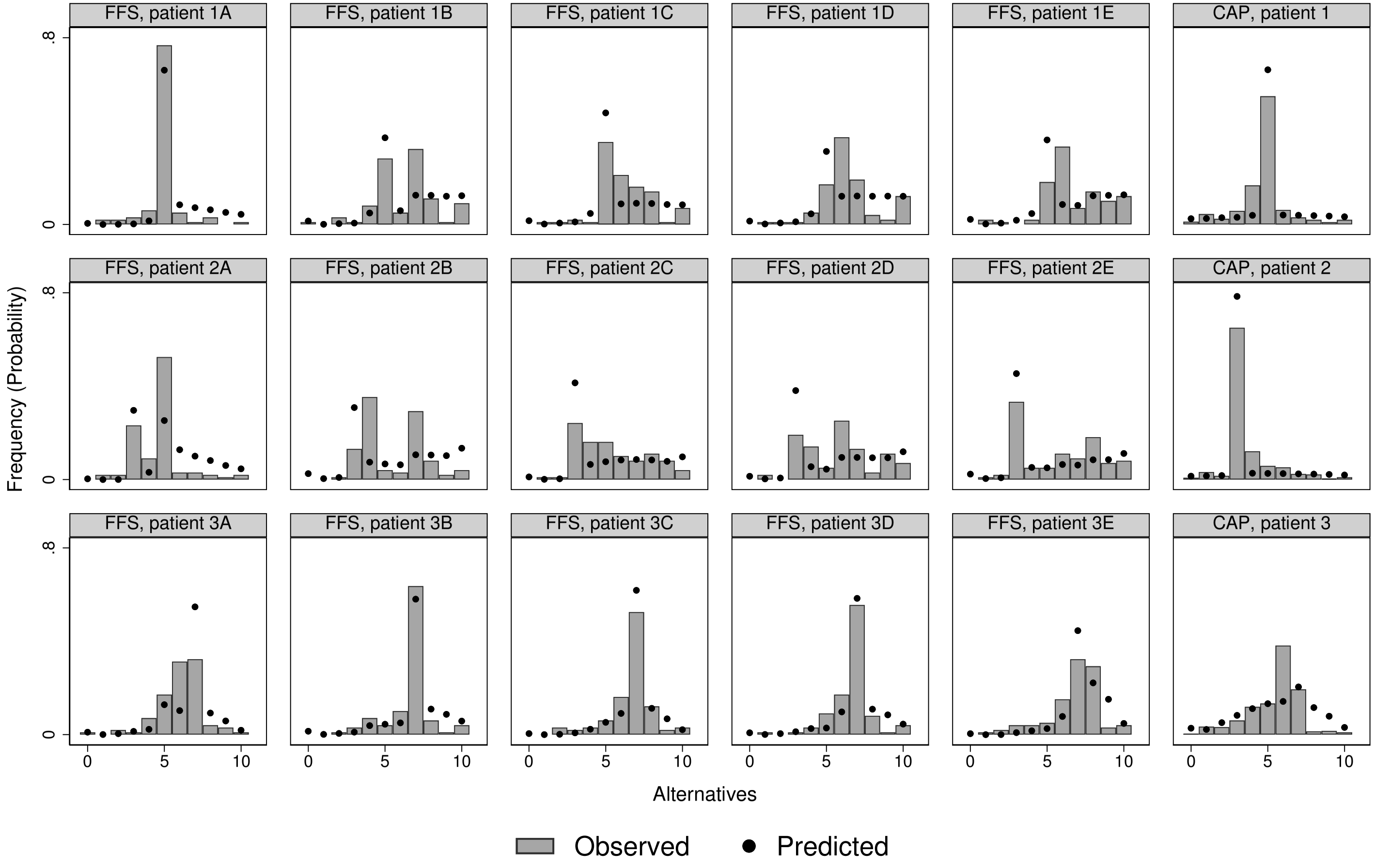
Based on the parameter estimates acquired from our data on student behavior in CAP and FFS, and doctor behavior in CAP only, we can predict quite closely the behavior of medical doctors under FFS. Vice versa, we can predict quite closely the behavior of medical doctors under CAP when information on doctor behavior in CAP is excluded from our estimation.

Empirical support is provided in Figure 3, where the observed and predicted behavior of Chinese medical doctors in FFS and CAP is shown. There are in total 198 unique treatment alternatives in the experiment, 165 treatment alternatives for the 15 different choice scenarios in FFS, and 33 treatment alternatives for the three different choice scenarios in CAP. For both FFS and CAP we apply statistical tests of matched pairs to test whether the observed frequency distribution differs from the predicted distribution. We cannot reject the null hypothesis that the observed and predicted frequencies for the alternative treatments in FFS and CAP, respectively, are the same (p-value=0.99 for both FFS and CAP, Fisher-Pitman permutation test for paired replicates). With reference to our third research question, we state:

RESULT 3: We find evidence that doctors' behavior in CAP (FFS) can be predicted by applying a subset of our data where doctors' behavior in CAP (FFS) is excluded.

Figure 3. Out of sample predictions of FFS and CAP behavior of Chinese doctors

Predictions based on estimates from a restricted version of (5), assuming $\alpha_n = \alpha \ \ \forall n$. Estimates acquired while excluding FFS (CAP) choices of Chinese doctors from estimation.



6. Discussion and concluding remarks.

We introduce a fully incentivized laboratory experiment, which extends the well-known experiment of Hennig-Schmidt et al. (2011) by including two payment schemes and twice the number of individual-level observations. We broaden the set of subject pools by recruiting Chinese medical doctors and medical students. Our results replicate the findings of Hennig-Schmidt et al. (2011). They corroborate the general results in the health economics literature that FFS payment encourages higher service volumes than CAP. Volumes under FFS are in general higher than what is optimal for the patient if this is profitable for the provider, and vice versa for CAP systems. Interestingly, we observe one instance in which volumes are higher than optimal for the patient even under CAP. While culture- and country-specific effects are found in many experimental studies, we provide an example where the same qualitative response to experimental conditions is observed in three different subject pools, from two different countries.

Our results suggest that preferences of subjects from very different subject pools are similar, and that the small differences in observed behavior across subject pools can be attributed to between-group-differences in the degree of randomness in behavior. Failure to reject a hypothesis could, however, be the result of insufficient statistical power. Given the small to moderate sample sizes of the the three subject pools, we can of course not rule out the possibility that small differences in patient-regarding preferences exist. We note, however, that the differences in point estimates between the three groups are relatively small. Further, there is evidence that one can provide accurate out-of-sample predictions of doctor behavior under conditions for which only data on students' behavior are available, implying that experimental behavioral data can provide valid knowledge, which reaches beyond the subject pools under study. Failure to reject a hypothesis could also result from model misspecification. Conclusions are robust to several alternative model specifications. In particular, we have estimated a mixed logit model with normal distributed coefficients, allowing for (mean) preference parameters to differ across subject pools, and allowing for the scale parameter to depend on experience, see Supplementary Material, Section D. When we compare means of subjects' marginal rates of substitution across subject pools (Hole, 2007), the conclusion remains the same.

One of the basic features of our experiment is that the monetary equivalent of the patient benefit is beneficial for real patients outside the laboratory in order to provide an incentive for participants to care for the patient when making their decisions. Hennig-Schmidt et al. (2011) transferred this money to the Christoffel Blindenmission, a German charity, which used the monetary transfers exclusively to support surgical treatments of cataract patients. We, however, chose to transfer the money to the in-hospital-account of one real patient of Shandong University Cancer Hospital to be used exclusively for his cancer treatment. It has been argued that this modified procedure weakens the medical context of the experiment compared to the original procedure and induces the participants to take the medical framing less seriously. Our motivation for modifying the protocol was to account for important cultural differences between China and Germany regarding the participants' possibly negative perception of the credibility of charitable organizations, and due to the fact that a charity similar to Christoffel Blindenmission did not exist in China. We argue that the medical frame in the present experiment is unlikely to be less salient than in the German experiment. Having the monitor observe that the money was paid into a patient's in-hospital-account provides credibility to the procedure. Moreover, the participants of our experiment were familiar with the Chinese health insurance system that captures only basic needs (Meng et al, 2019). They were aware that patients had to bear rather high out-of-pocket payments, which at the time of our experiment amounted to around 35% of their total treatment costs (Fang et al. 2019) – resulting in rather high own payments when a patient suffered from cancer. In total, 19,814 RMB (€2,377.68), about four times the average monthly wage of about 4,650 RMB in Jinan in 2012 (China Statistics Press, 2013), was transferred to the patient's in-hospital account. Given the relatively high level of observed patient benefit in the experiment, a plausible interpretation is that the participants took the medical framing seriously.

In our analysis, we assume individuals are boundedly rational. An interesting question is how individuals would have behaved if they were perfectly rational, such that the influence of noise in the optimization was absent, i.e., $\sigma_{nt} = 0$. We investigate how behavior in the experiment would have been under the perfect rationality assumption. We find that whether humans are regarded as perfectly rational or boundedly rational, has a substantial influence on the predicted behavioral response to a payment reform (see the aggregate quantities of service provision under the assumption of perfect rationality for FFS and CAP in Table B2 in the Supplementary Material). In the case of our chosen experimental parameters, the predicted difference in behavior between the two payment schemes is exaggerated if one assumes perfectly rational individuals. Implications for policy-making can be that behavioral predictions are distorted as the following example illustrates: Imagine a policy-maker who is in favor of replacing a FFS scheme by a CAP scheme if the CAP scheme was expected to reduce average service quantity for patients by only 1.6 units. This policy-maker might well prefer to prolong the FFS scheme if behavioral responses become too strong, and a quantity reduction of 2.5 units was to be expected.

The literature on how financial incentives affect behavior is vast, and has enabled the development of evidence-based policies. An example is the use of payment schemes that combines fixed and activity-based payment in health sectors in many countries. Given the existence of theory and empirical methods for analyzing behavior under the bounded rationality assumption, it is surprisingly little economic research that addresses the important question on how observable variables affect the degree of rational decision-making, under given financial incentives. If humans are boundedly rational, regulators cannot expect that even a perfectly designed payment scheme will result in optimal decisions, at least not all the time. More scientific knowledge must be acquired in order to develop and implement policies that improve the quality of decision-making. The case of medical decisions is one of many examples where decision-making quality is expected to affect welfare.

The frequency of failed replication attempts, for example by Camerer et al. (2016) and Camerer et al. (2018), has caused much debate in the scientific community, and hot topics include the causes of replication failure and how future research should adapt in order to promote scientific excellence. Some researchers, e.g. Shrout and Rodgers (2018), propose that replication failures are caused by inappropriate power analysis, and Benjamin et al. (2018) propose to simply strengthen the requirement for statistical significant results and requiring conclusions to be drawn on lower p-values. Some argue that publication bias contributes to the likelihood of replication failure, and Andrews and Kasy (2019) even propose methods to adjust for publication bias when conducting meta-analysis. Loken and Gelman (2017) provide a reminder on fundamental aspects of scientific research: It is not plausible to assume that noise and measurement errors are absent, even when data are from controlled experiments. We prominently acknowledge the presence of noise in this study. We show that the influence of random noise on choice is significantly lower when subjects become experienced.

To the best of our knowledge, none of the contributions that report results from replicated experiments discusses or describes how much experimental experience their participating subjects have compared to the subjects in the original experiments. When replication attempts are conducted later in history, and at different geographical locations, it is not obvious that the recruited subjects will be equally experienced with participating in experiments. An interesting research question is whether replication failures are caused by differences in experimental subjects' experience with participating in laboratory experiments, and if so, whether experience-effects depend on geographical location and type of experiment.

Acknowledgements.

Comments, questions and suggestions from two anonymous reviewers have contributed to substantial improvements of this paper. We thank Lin Jing for his assistance in conducting experiments. We are grateful to Chaoliang Yang for translating all the material to Chinese, and for assistance in conducting the experiments. We are grateful to Jingyi Luo for translating material from Chinese to English. We have benefited from discussions with Daniel Wiesen, participants at the IHEA-World Congress in Sydney in 2013, ASHECON 2014, the Second Workshop in Behavioral and Experimental Health Economics in Hamilton, Canada, 2015, and The Nordic Health Economists Study Group Meeting in Reykjavik, Iceland 2019. Thanks to the authors of Hennig-Schmidt et al. (2011) for providing us with their data and experimental design. We are grateful for financial support from the Independent Innovation Fund of Shandong University (grant no. 2012JC038) and the National Natural Science Foundation of China (grant no. 71373146) for funding the experiments conducted at Shandong University, and from the Research Council of Norway (Project-No. 231776) for support to Geir Godager, Heike Hennig-Schmidt and Tor Iversen. Part of Heike Hennig-Schmidt's research for this article was done within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100.

Literature

Afriat, S. N. (1972): "Efficiency estimation of production functions," *International Economic Review*, 13, 568–598.

——— (1973): "On a system of inequalities in demand analysis: an extension of the classical method," *International Economic Review*, 14, 460–472.

Allard, M., I. Jelovac, and P. Léger (2011): "Treatment and Referral Decisions under Different Physician Payment Mechanisms," *Journal of Health Economics*, 30, 880–893.

Andreoni, J. and J. Miller (2002): "Giving according to GARP: An experimental test of the consistency of preferences for altruism," *Econometrica*, 70, 737–753.

Andrews, I. and M. Kasy (2019): "Identification of and correction for publication bias," *American Economic Review*, 109, 2766–94.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. (2018): "Redefine statistical significance," *Nature Human Behaviour*, 2, 6.

Blundell, R. (2005): "How revealing is revealed preference?" *Journal of the European Economic Association*, 3, 211–235.

Brosig-Koch, J., H. Hennig-Schmidt, N. Kairies-Schwarz, and D. Wiesen (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.

——— (2017): "The effects of introducing mixed payment systems for physicians: Experimental evidence," *Health Economics*, 26, 243–262.

Brosig-Koch, J., H. Hennig-Schmidt, J. Kokot, N. Kairies-Schwarz, and D. Wiesen (2019): "Physician performance pay: Experimental evidence." *Discussion paper, SSRN-id3467583.*

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. (2016): "Evaluating replicability of laboratory experiments in economics," *Science*, 351, 1433–1436.

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. (2018): "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, 2, 637.

China Statistics Press (2013): "Jinan Statistical Yearbook 2013," http://cdi.cnki.net/Search/ReportPreview?FileName=N2013110084000110, accessed: 2019-11-26.

Choi, S., S. Kariv, W. Müller, and D. Silverman (2014): "Who is (more) rational?" *American Economic Review*, 104, 1518–50.

Chone, P. and C. A. Ma (2011): "Optimal Health Care Contract under Physician Agency," *Annales d'Economie et de Statistique*, 101/202, 229–256.

Clemens, J. and J. D. Gottlieb (2014): "Do physicians' financial incentives affect medical treatment and patient health?" *American Economic Review*, 104, 1320–49.

Costa, A., A. Foucart, I. Arnon, M. Aparici, and J. Apesteguia (2014): ""Piensa" twice: On the foreign language effect in decision making," *Cognition*, 130, 236–254.

Cox, J. C. (1997): "On testing the utility hypothesis," *The Economic Journal*, 107, 1054–1078.

Cox, J. C., E. Green, and H. Hennig-Schmidt (2016): "Experimental and behavioral economics of healthcare," *Journal of Economic Behavior & Organization*, 131, A1–A4.

Dagsvik, J. K. (2008): "Axiomatization of stochastic models for choice under uncertainty," *Mathematical Social Sciences*, 55, 341–370.

Dagsvik, J. K. and S. R. Hoff (2011): "Justification of functional form assumptions in structural models: applications and testing of qualitative measurement axioms," *Theory and Decision*, 70, 215–254.

Devlina, R. and S. Sarma (2008): "Do Physician Remuneration Schemes Matter? The Case of Canadian Family Physicians," *Journal of Health Economics*, 27, 1168–1181.

Eckel, C. and P. Grossman (1996): "Altruism in Anonymous Dictator Games," *Games and Economic Behavior*, 16, 181–191.

Eggleston, K. (2012): "Health care for 1.3 billion: China's remarkable work in progress," *Milken Institute Review*, 16–27.

Ellis, R. P. and T. G. McGuire (1986): "Provider Behavior under Prospective Reimbursement: Cost Sharing and Supply," *Journal of Health Economics*, 5, 129–151.

——— (1990): "Optimal Payment Systems for Health Services," *Journal of Health Economics*, 9, 375–396.

Falk, A. and E. Fehr (2003): "Why labour market experiments?" *Labour Economics*, 10, 399–406.

Falk, A. and J. Heckman (2009): "Lab experiments are a major source of knowledge in the social sciences," *Science*, 326, 535–538.

Fang, H., K. Eggleston, K. Hanson, and M. Wu (2019): "Enhancing financial protection under China's social health insurance to achieve universal health coverage," *British Medical Journal*, 365, l2378.

Février, P. and M. Visser (2004): "A study of consumer behavior using laboratory data," *Experimental Economics*, 7, 93–114.

Fiebig, D. G., M. P. Keane, J. Louviere, and N. Wasi (2010): "The generalized multinomial logit model: accounting for scale and coefficient heterogeneity," *Marketing Science*, 29, 393–421.

Fink, C. and N. Kairies-Schwarz (2019): "Performance Pay in Hospitals: An Experiment on Bonus-Malus Incentives," *mimeo, University of Duisburg-Essen.*

Fischbacher, U. (2007): "Z-tree: Zurich Toolboox for Readymade Economic Experiments – Experimenter's Manual," *Experimental Economics*, 10, 171–178.

Gächter, S. and J. F. Schulz (2016): "Intrinsic honesty and the prevalence of rule violations across societies," *Nature*, 531, 496.

Gaynor, M. and P. Gertler (1995): "Moral Hazard and Risk Spreading in Partnerships," *Rand Journal of Economics*, 26, 591–613.

Ge, G., G. Godager, and J. Wang (2019): "Do physicians care about patients' utility? Evidence from an experimental study of treatment choices under demand-side cost sharing," *HERO Working Paper Series*, 2019:2.

Glimcher, P. W. (2011): *Foundations of neuroeconomic analysis*, Oxford: Oxford: Oxford University Press.

Godager, G., H. Hennig-Schmidt, and T. Iversen (2016): "Does performance disclosure influence physicians' medical decisions? An experimental study," *Journal of Economic Behavior & Organization*, 131, 36–46.

Godager, G. and D. Wiesen (2013): "Profit or Patients' Health Benefit? Exploring the Heterogeneity in Physician Altruism," *Journal of Health Economics*, 32, 1105–116.

Gosden, T., F. Forland, I. Kristiansen, M. Sutton, B. Leese, A. Guiffrida, M. Sergison, and L. Pedersen (2001): "Impact of Payment Method on Behavior of Primary Care Physicians: A Systematic Review," *Journal of Health Services Research and Policy*, 6, 44–54.

Gu, Y., A. R. Hole, and S. Knox (2013): "Fitting the generalized multinomial logit model in Stata," *Stata Journal*, 13, 382–397.

Hafner, L., S. Reif, and M. Seebauer (2017): "Physician Behavior under Prospective Payment Schemes: Evidence from artefactual field and lab experiments," *FAU Discussion Papers in Economics No. 18-2017.*

Hands, D. W. (2013): "Foundations of contemporary revealed preference theory," *Erkenntnis*, 78, 1081–1108.

——— (2014): "Paul Samuelson and revealed preference theory," *History of Political Economy*, 46, 85–116.

Hennig-Schmidt, H., R. Selten, and D. Wiesen (2011): "How Payment Systems Affect Physicians' Provision Behavior – An Experimental Investigation," *Journal of Health Economics*, 30, 637–646.

Hennig-Schmidt, H. and D. Wiesen (2014): "Other-regarding behavior and motivation in health care provision: An experiment with medical and non-medical students," *Social Science & Medicine*, 108, 156–165.

Herrmann, B., C. Thöni, and S. Gächter (2008): "Antisocial punishment across societies," *Science*, 319, 1362–1367.

Hess, S. and J. M. Rose (2012): "Can scale and coefficient heterogeneity be separated in random coefficients models?" *Transportation*, 39, 1225–1239.

Hess, S. and K. Train (2017): "Correlation and scale in mixed logit models," *Journal of Choice Modelling*, 23, 1–8.

Hole, A. R. (2007): "A Comparison of Approaches to Estimating Confidence Intervals for Willingness to Pay Measures," *Health Economics*, 16, 827–840.

Holmes, T. P. and K. J. Boyle (2005): "Dynamic learning and context-dependence in sequential, attribute-based, stated-preference valuation questions," *Land Economics*, 81, 114–126.

Iversen, T. (2004): "The effects of a patient shortage on general practitioners' future income and list of patients," *Journal of Health Economics*, 23, 673–694.

Iversen, T. and H. Lurås (2000): "The Effect of Capitation on GPs' Referal Decision," *Health Economics*, 9, 199–210.

Kahneman and A. Tversky (1979): "An analysis of decision under risk," *Econometrica*, 47, 263–292.

Keser, C., C. Montmarquette, M. Schmidt, and C. Schnitzler (2014): "Custom-made healthcare–An experimental investigation," *cege Discussion Papers.*

Li, J. (2018): "Plastic surgery or primary care? Altruistic preferences and expected specialty choice of US medical students," *Journal of Health Economics*, 62, 45–59.

Li, J., W. H. Dow, and S. Kariv (2017): "Social preferences of future physicians," *Proceedings of the National Academy of Sciences*, 114, E10291–E10300.

Lieberman, B. (1960): "Human behavior in a strictly determined 3×3 matrix game," *Behavioral Science*, 5, 317–322.

Loken, E. and A. Gelman (2017): "Measurement error and the replication crisis," *Science*, 355, 584–585.

Luce, R. D. (1959a): *Individual Choice Behavior a Theoretical Analysis*, Oxford, England: John Wiley.

——— (1959b): "On the possible psychophysical laws." *Psychological Review*, 66, 81.

Ma, C. A. (1994): "Health Care Payment Systems: Cost and Quality Incentives," *Journal of Economics and Management Strategy*, 3, 93–112.

Ma, C.-t. A. and H. Y. Mak (2019): "Incentives in Healthcare Payment Systems," in *Oxford Research Encyclopedia of Economics and Finance*, Oxford University Press.

Mattei, A. (2000): "Full-scale real tests of consumer behavior using experimental data," *Journal of Economic Behavior & Organization*, 43, 487–497.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. E. Zarembka, Academic Press, New York, 105–142.

——— (1981): "Econometric models of probabilistic choice," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden, MIT Press, 198–272.

——— (2001): "Economic Choices," *American Economic Review*, 91, 351–378.

McFadden, D., M. J. Machina, and J. Baron (1999): "Rationality for economists?" in *Elicitation of preferences*, ed. by F. B. and M. C.F., Springer, 73–110.

McGuire, T. G. (2000): "Physician Agency," in *Handbook of Health Economics, Vol. 1 A*, ed. by Cuyler and Newhouse, North-Holland, Amsterdam (The Netherlands), 461–536.

McKelvey, R. D. and T. R. Palfrey (1995): "Quantal response equilibria for normal form games," *Games and Economic*

Behavior, 10, 6–38.

Meng, Q., A. Mills, L. Wang, and Q. Han (2019): "What can we learn from China's health system reform?" *British Medical Journal*, 365, l2349.

Newhouse, J. P. (1996): "Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection," *Journal of Economic Literature*, 34, 1236–1263.

Olsen, S. B., J. Meyerhoff, M. R. Mørkbak, and O. Bonnichsen (2017): "The influence of time of day on decision fatigue in online food choice experiments," *British Food Journal*, 119, 497–510.

Samuelson, P. A. (1938): "A note on the pure theory of consumer's behaviour," *Economica*, 5, 61–71.

Scott, A., M. Liu, and J. Yong (2018): "Financial incentives to encourage value-based health care," *Medical Care Research and Review*, 75, 3–32.

Sen, A. (1973): "Behaviour and the Concept of Preference," *Economica*, 40, 241–259.

——— (1993): "Internal consistency of choice," *Econometrica*, 61, 495–521.

——— (1997): "Maximization and the Act of Choice," *Econometrica*, 65, 745–779.

Sen, A. K. (1977): "Rational fools: A critique of the behavioral foundations of economic theory," *Philosophy & Public Affairs*, 317–344.

Shrout, P. E. and J. L. Rodgers (2018): "Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis," *Annual Review of Psychology*, 69, 487–510, pMID: 29300688.

Simon, H. A. (1957): *Models of man; social and rational.*, Wiley, New York.

——— (1979): "Rational decision making in business organizations," *The American Economic Review*, 69, 493–513.

Sørensen, R. and J. Grytten (2003): "Service Production and Contract Choice in Primary Physician Services," *Health Policy*, 66, 73–93.

Thurstone, L. L. (1927a): "A law of comparative judgment." *Psychological Review*, 34, 273.

——— (1927b): "Psychophysical analysis," *The American Journal of Psychology*, 38, 368–389.

Train, K. E. (2009): *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge (UK).

Tversky, A. (1972): "Choice by elimination," *Journal of Mathematical Psychology*, 9, 341–367.

Tversky, A. and D. Kahneman (1974): "Judgment under uncertainty: Heuristics and biases," *Science*, 185, 1124–1131.

Van Dijk, C. E., B. Van den Berg, R. A. Verheij, P. Spreeuwenberg, P. P. Groenewegen, and D. H. De Bakker (2013): "Moral hazard and supplier-induced demand: Empirical evidence in general practice," *Health Economics*, 22, 340–352.

Varian, H. R. (1982): "The nonparametric approach to demand analysis," *Econometrica*, 50, 945–973.

——— (1983): "Non-parametric tests of consumer behaviour," *The Review of Economic Studies*, 50, 99–110.

——— (1991): *Goodness-of-fit for revealed preference tests*, Department of Economics, University of Michigan Ann Arbor.

Yip, W. and W. C. Hsiao (2008): "The Chinese health system at a crossroads," *Health Affairs*, 27, 460–468.

Yip, W. C.-M., W. Hsiao, Q. Meng, W. Chen, and X. Sun (2010): "Realignment of incentives for health-care providers in China," *The Lancet*, 375, 1120–1130.

Appendix A. Experimental parameters

Table A1: Experimental parameters

	Payment	Var	0	1	2	3	4	5	6	7	8	9	10
I	FFS	R_{jA}(q)	0.00	1.70	3.40	5.10	5.80	10.50	11.00	12.10	13.50	14.90	16.60
		$R_{jB}(q)$	0.00	1.00	2.40	3.50	8.00	8.40	9.40	16.00	18.00	20.00	22.50
		$R_{jC}(q)$	0.00	1.80	3.60	5.40	7.20	9.00	10.80	12.60	14.40	16.20	18.30
		$R_{jD}(q)$	0.00	2.00	4.00	6.00	8.00	8.00	15.00	16.90	18.90	21.30	23.60
		$R_{jE}(q)$	0.00	1.00	2.00	6.00	6.70	7.60	11.00	12.30	18.00	20.50	23.00
	CAP	$R(q)$	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00
II	FFS,CAP	c(q)	0.00	0.10	0.40	0.90	1.60	2.50	3.60	4.90	6.40	8.10	10.00
III	FFS	$\pi_{jA}(q)$	0.00	1.60	3.00	4.20	4.20	8.00	7.40	7.20	7.10	6.80	6.60
		$\pi_{jB}(q)$	0.00	0.90	2.00	2.60	6.40	5.90	5.80	11.10	11.60	11.90	12.50
		$\pi_{jC}(q)$	0.00	1.70	3.20	4.50	5.60	6.50	7.20	7.70	8.00	8.10	8.30
		$\pi_{jD}(q)$	0.00	1.90	3.60	5.10	6.40	5.50	11.40	12.00	12.50	13.20	13.60
		$\pi_{jE}(q)$	0.00	0.90	1.60	5.10	5.10	5.10	7.40	7.40	11.60	12.40	13.00
	CAP	$\pi(q)$	12.00	11.90	11.60	11.10	10.40	9.50	8.40	7.10	5.60	3.90	2.00
IV	FFS,CAP	B1k(q)	0.00	0.75	1.50	2.00	7.00	10.00	9.50	9.00	8.50	8.00	7.50
		$B_{2k}(q)$	0.00	1.00	1.50	10.00	9.50	9.00	8.50	8.00	7.50	7.00	6.50
		$B_{3k}(q)$	0.00	0.75	2.20	4.05	6.00	7.75	9.00	9.45	8.80	6.75	3.00

Note: This table shows all experimental parameters. $R_{jk}(q)$ denotes physicians' payment for patient type j and illness k. Under FFS, $R_{jk}(q)$ varies with illnesses k and increases in q, whereas under CAP, $R_{jk}(q)$ remains constant. The costs for providing medical services $c_{jk}(q)$ increase in q and are the same under all experimental conditions. The physicians' profit $\pi_{jk}(q)$ is equal to $R_{jk}(q) - c_{jk}(q)$. $B_{jk}(q)$ denotes the patient benefit for the three patient types $j = 1, 2, 3$ held constant across conditions.

Are patient-regarding preferences stable?

Evidence from a laboratory experiment with physicians and medical students from different countries

Jian Wang[1 2], Tor Iversen[2], Heike Hennig-Schmidt[2 3 4], Geir Godager[2 5]

Supplementary Material*

[1] Dong Fureng Institute of Economic and Social Development, Wuhan University, China

[2] Institute of Health and Society, Department of Health Management and Health Economics, University of Oslo, Norway

[3] BonnEconLab, University of Bonn, Germany

[4] National Research University Higher School of Economics (HSE), Moscow, Russian Federation

[5] Health Services Research Unit, Akershus University Hospital, Norway

∗ This is supplementary material for a forthcoming article in *European Economic Review*.

A. Experimental parameters

Table A1: Experimental parameters

	Payment	Var	0	1	2	3	4	5	6	7	8	9	10
I	FFS	R_{jA}(q)	0.00	1.70	3.40	5.10	5.80	10.50	11.00	12.10	13.50	14.90	16.60
		$R_{jB}(q)$	0.00	1.00	2.40	3.50	8.00	8.40	9.40	16.00	18.00	20.00	22.50
		$R_{jC}(q)$	0.00	1.80	3.60	5.40	7.20	9.00	10.80	12.60	14.40	16.20	18.30
		$R_{jD}(q)$	0.00	2.00	4.00	6.00	8.00	8.00	15.00	16.90	18.90	21.30	23.60
		$R_{jE}(q)$	0.00	1.00	2.00	6.00	6.70	7.60	11.00	12.30	18.00	20.50	23.00
	CAP	$R(q)$	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00
II	FFS,CAP	c(q)	0.00	0.10	0.40	0.90	1.60	2.50	3.60	4.90	6.40	8.10	10.00
III	FFS	$\pi_{jA}(q)$	0.00	1.60	3.00	4.20	4.20	8.00	7.40	7.20	7.10	6.80	6.60
		$\pi_{jB(q)}$	0.00	0.90	2.00	2.60	6.40	5.90	5.80	11.10	11.60	11.90	12.50
		$\pi_{jC(q)}$	0.00	1.70	3.20	4.50	5.60	6.50	7.20	7.70	8.00	8.10	8.30
		$\pi_{jD(q)}$	0.00	1.90	3.60	5.10	6.40	5.50	11.40	12.00	12.50	13.20	13.60
		$\pi_{jE(q)}$	0.00	0.90	1.60	5.10	5.10	5.10	7.40	7.40	11.60	12.40	13.00
	CAP	$\pi(q)$	12.00	11.90	11.60	11.10	10.40	9.50	8.40	7.10	5.60	3.90	2.00
IV	FFS,CAP	B1k(q)	0.00	0.75	1.50	2.00	7.00	10.00	9.50	9.00	8.50	8.00	7.50
		$B_{2k}(q)$	0.00	1.00	1.50	10.00	9.50	9.00	8.50	8.00	7.50	7.00	6.50
		$B_{3k}(q)$	0.00	0.75	2.20	4.05	6.00	7.75	9.00	9.45	8.80	6.75	3.00

Note: This table shows all experimental parameters. $R_{jk}(q)$ denotes physicians' payment for patient type j and illness k. Under FFS, $R_{jk}(q)$ varies with illnesses k and increases in q, whereas under CAP, $R_{jk}(q)$ remains constant. The costs for providing medical services $c_{jk}(q)$ increase in q and are the same under all experimental conditions. The physicians' profit $\pi_{jk}(q)$ is equal to $R_{jk}(q) - c_{jk}(q)$. $B_{jk}(q)$ denotes the patient benefit for the three patient types $j = 1, 2, 3$ held constant across conditions.

B. Additional analyses. The effect of payment scheme

Hennig-Schmidt et al.'s (2011) original question about the effect of payment schemes on the quantity of care, level of patient benefit, and the probability of maximizing patient benefit can be assessed jointly in ordered regression models that account for correlation between observations of the same individual. We present such an empirical analysis in table B0.

Table B0. Results from ordered logistic regression (q and $B(q)$) and logistic regression (maximized $B(q)$ yes/no) with individual specific random effects

	q				$B(q)$				maximized $B(q)$			
	OR		[95% Conf. Int]†		OR		[95% Conf. Int]†		OR		[95% Conf. Int]†	
FFS	11.66	***	8.25	16.48	0.73	*	0.52	1.02	0.44	***	0.31	0.64
Doctor	0.98		0.65	1.48	0.62	**	0.43	0.89	0.51	***	0.34	0.75
German	1.17		0.72	1.90	0.98		0.58	1.67	1.12		0.61	2.07
Decision#												
2	3.26	***	2.53	4.20	0.16	***	0.10	0.26	0.15	***	0.10	0.24
3	2.77	***	2.18	3.52	0.25	***	0.16	0.38	0.16	***	0.11	0.24
4	3.83	***	3.01	4.87	0.24	***	0.15	0.38	0.13	***	0.09	0.21
5	4.77	***	3.52	6.48	0.20	***	0.12	0.33	0.15	***	0.10	0.23
6	0.29	***	0.22	0.38	0.25	***	0.15	0.42	0.15	***	0.10	0.24
7	0.80		0.56	1.16	0.24	***	0.13	0.42	0.15	***	0.09	0.24
8	0.52	***	0.36	0.75	0.31	***	0.17	0.54	0.26	***	0.16	0.40
9	0.86		0.58	1.27	0.20	***	0.11	0.37	0.20	***	0.13	0.32
10	0.68		0.42	1.10	0.33	***	0.18	0.63	0.41	***	0.26	0.63
11	5.76	***	4.36	7.62	0.05	***	0.03	0.08	0.10	***	0.07	0.15
12	8.22	***	6.21	10.87	0.06	***	0.04	0.10	0.23	***	0.16	0.33
13	9.62	***	7.29	12.70	0.06	***	0.04	0.10	0.18	***	0.12	0.25
14	10.19	***	7.69	13.49	0.06	***	0.04	0.10	0.22	***	0.15	0.32
15	12.58	***	9.14	17.32	0.05	***	0.03	0.08	0.07	***	0.05	0.11
Log pseudolikelihood	-7862.3416				-10086.548				-2689.0698			
# obs =	4785											
# subjects =	319											
# decisions =	15 (decisions in the first part of the session, $t = 1$)											

† Confidence intervals are based on robust standard errors.

Table B1. The prevalence of choice variation when incentive variation is absent in CAP

	Subsample			
Patient 1	All	Chinese	Chinese	German
# unique actions		student	doc	student
1 (No variation)	146	99	34	13
2	73	47	22	4
3	41	20	18	3
4	23	7	15	1
5	16	5	10	1
Total	299	178	99	22
Patient 2	All	Chinese	Chinese	German
# unique actions		student	doc	student
1 (No variation)	184	128	43	13
2	56	29	22	5
3	27	6	18	3
4	25	13	11	1
5	7	2	5	0
Total	299	178	99	22
Patient 3	All	Chinese	Chinese	German
# unique actions		student	doc	student
1 (No variation)	113	67	36	10
2	110	70	31	9
3	46	31	14	1
4	18	7	9	2
5	12	3	9	0
Total	299	178	99	22

This table shows the frequency of choice variation when subjects make 5 repeated treatment choices for the same patients (1 ,2 and 3) under CAP. Sample: 178 Chinese students, 99 Chinese doctors and 22 German students.

Table B2. Predicted behavior of Chinese doctors and medical students under the perfect rationality assumption in CAP and FFS, assuming $\sigma = 0$.

Payment system	Doctors Quantity	Medical students Quantity	Total Quantity
CAP	4.33	4.67	4.55
FFS	7.33	6.87	7.03

This table describes how aggregate quantities of service provision over payment systems and subject pools would have appeared if randomness in decision-making was absent, $\sigma = 0$.

Literature

Hennig-Schmidt, H., R. Selten, and D. Wiesen (2011): "How Payment Systems Affect Physicians' Provision Behavior – An Experimental Investigation," *Journal of Health Economics*, 30, 637–646.

C. Experiment material

C1: Instructions of the experiment

> [Numbers/text in brackets refer to the conditions where doctors participate.]
>
> {Sentences/decision screens in braces are inserted into the instructions either in condition FFS or in condition CAP.}

Instructions Part 1
General Information

In the following experiment, you will make a couple of decisions. Following the instructions and depending on your decisions, you can earn money. It is therefore very important that you read the instructions carefully.

You take your decisions anonymously on your computer screen. During the experiment, you are not allowed to talk to any other participant. Whenever you have a question, please raise your hand. The experimenter will answer your question in private in your cubicle. If you disregard these rules, you can be excluded from the experiment without receiving any payment. All amounts of money in the experiment are stated in Token. At the end of the experiment, your earnings will be converted into RMB at an exchange rate of 10 Token = 1 [6] RMB and paid to you in cash.

The experiment consists of two parts. We we will inform you now on the decision situation in Part 1. We will provide you with the instructions for Part 2 as soon as Part 1 has ended. Please note that your decisions in Part 1 have no influence on your decisions in Part 2 and vice versa.

Your decisions in Part 1 of the experiment
During the experiment, you are in the role of a physician. You have to make 15 decisions regarding the treatment of patients. All participants of this experiment take their decisions in the role of physicians. You decide on the quantity of medical services you want to provide for given clinical symptoms of a patient.

You decide on your computer screen where five different kinds of clinical symptoms – A, B, C, D, and E – of three different patient types – 1, 2, and 3 – will be shown one after another. For each patient you can provide 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Your remuneration is as follows:

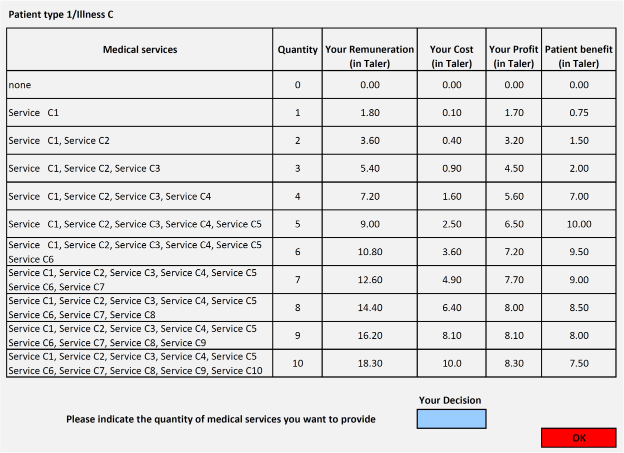
Condition CAP: For each patient you receive a lump-sum payment that is independent of the quantity of medical services.

Condition FFS: A different payment is assigned to each quantity of medical services. The payment increases in the quantity of medical services.

While deciding on the quantity of medical services, in addition to your payment you determine the costs you incur when providing these services. Costs increase with increasing quantity provided. Your profit in Token is calculated by subtracting your costs from your payment.

A certain benefit for the patient is assigned to each quantity of medical services, the patient benefit that the patient gains from your provision of services (treatment). Therefore, your decision on the quantity of medical services not only determines your own profit, but also the patient benefit. An example for a decision situation is given on the following screen.

{Decision screen for patient 1C under FFS and CAP}

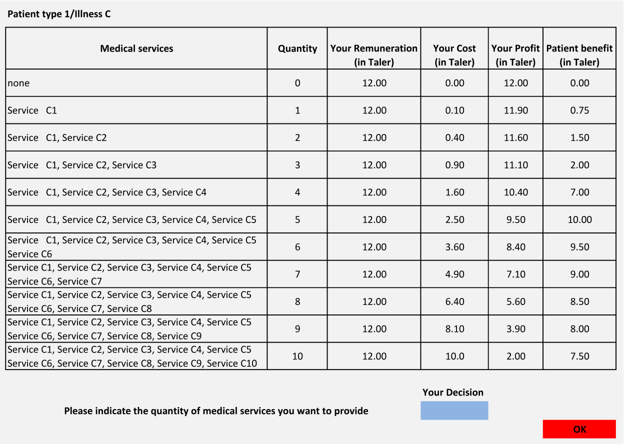


Patient type 1/Illness C

Medical services	Quantity	Your Remuneration (in Taler)	Your Cost (in Taler)	Your Profit (in Taler)	Patient benefit (in Taler)
none	0	0.00	0.00	0.00	0.00
Service C1	1	1.80	0.10	1.70	0.75
Service C1, Service C2	2	3.60	0.40	3.20	1.50
Service C1, Service C2, Service C3	3	5.40	0.90	4.50	2.00
Service C1, Service C2, Service C3, Service C4	4	7.20	1.60	5.60	7.00
Service C1, Service C2, Service C3, Service C4, Service C5	5	9.00	2.50	6.50	10.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6	6	10.80	3.60	7.20	9.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7	7	12.60	4.90	7.70	9.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8	8	14.40	6.40	8.00	8.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9	9	16.20	8.10	8.10	8.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9, Service C10	10	18.30	10.0	8.30	7.50

Your Decision

Please indicate the quantity of medical services you want to provide



Patient type 1/Illness C

Medical services	Quantity	Your Remuneration (in Taler)	Your Cost (in Taler)	Your Profit (in Taler)	Patient benefit (in Taler)
none	0	12.00	0.00	12.00	0.00
Service C1	1	12.00	0.10	11.90	0.75
Service C1, Service C2	2	12.00	0.40	11.60	1.50
Service C1, Service C2, Service C3	3	12.00	0.90	11.10	2.00
Service C1, Service C2, Service C3, Service C4	4	12.00	1.60	10.40	7.00
Service C1, Service C2, Service C3, Service C4, Service C5	5	12.00	2.50	9.50	10.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6	6	12.00	3.60	8.40	9.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7	7	12.00	4.90	7.10	9.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8	8	12.00	6.40	5.60	8.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9	9	12.00	8.10	3.90	8.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9, Service C10	10	12.00	10.0	2.00	7.50

Your Decision

Please indicate the quantity of medical services you want to provide

You decide on the quantity of medical services on your computer screen by typing an integer between 0 and 10 into the box labeled ”Your Decision”.

After all participants have taken their decisions for the respective patient you will proceed to the next patient. There are no real, but abstract patients participating in this experiment. Yet, the patient benefit, which an abstract patient receives by your providing medical services, will be beneficial for a real patient. The total amount of patient benefit determined by your 15 decisions will be provided to a patient with cancer treated in Shandong Qilu Hospital [Shandong University Cancer Hospital]. The money will be directly transferred to the patient's in-hospital account to finance part of his/her treatment fee.

Each time you make a decision on the quantity of medical services you will be informed on your profit and the patient benefit. After you have made your 15 decisions in Part 1 of the experiment you will get to know your total profit and the corresponding total patient benefit.

Earnings in Part 1 of the experiment

After you have made your decisions in Part 1 of the experiment, your overall earnings will be calculated by summing up your profits from providing medical services to the 15 patients. This amount will be converted from Token into RMB. Your earnings of Part 1 of the experiment together with the earnings of Part 2 will be paid to you in cash at the end of the experiment (rounded to 1 Yuan).

The patient benefit gained by all 15 patients will be converted into RMB at the end of the experiment, too, and will be transferred to the real patient's in-hospital account. To this end the experimenter and a monitor will go together to Shandong Qilu Hospital [Shandong University Cancer Hospital]. After the transfer, the signed receipt will be scanned into electronic form and will be sent to all the participants via e-mail in order to ensure the authenticity of the above process. Personal information will be blinded black to respect the patient's privacy.

After the end of Part 2 of the experiment, one participant is randomly assigned the role of the monitor. The monitor receives a payment of 50 [200] RMB in addition to the payment from the experiment. In the end, the monitor signs a form to verify that the procedure described above was actually carried out. This form will be sent to all participants together with the receipt via e-mail.

Next, please answer some questions familiarizing you with the decision situation. After your 15 decisions, please answer some further questions on your screen.

Instructions Part 2

The experiment will now be repeated including one change. Like in Part 1 you will make 15 decisions. After these 15 decisions the experiment will end.

The General Information from Part 1 also applies for Part 2 of the experiment.

Your decisions in Part 2 of the experiment

Also in Part 2 of the experiment, you are in the role of a physician and you have to make 15 decisions regarding the treatment of patients. All participants take their decisions in the role of physicians. You decide on the quantity of medical services you want to provide for given clinical symptoms of a patient.

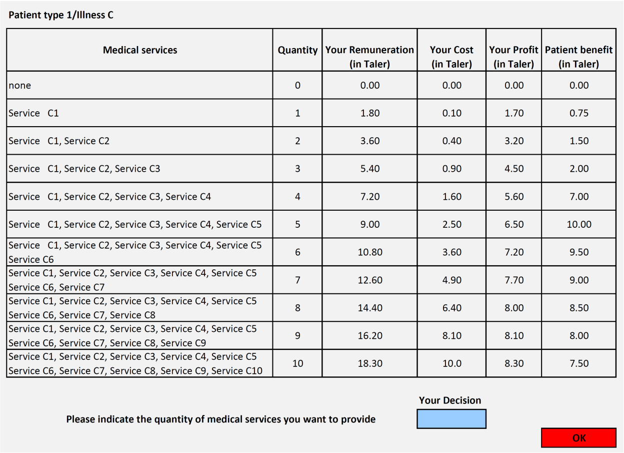
Like in Part 1 you decide on your computer screen where five different kinds of clinical symptoms A, B, C, D, and E of three different patient types (1, 2, and 3) will be shown one after another. For each patient you can provide 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Your remuneration is as follows:
{Condition CAP: For each patient you receive a lump-sum payment that is independent of the quantity of medical services.}
{Condition FFS: A different payment is assigned to each quantity of medical services. The payment increases in the quantity of medical services.}

As in Part 1, while deciding on the quantity of medical services, in addition to your payment you determine the costs you incur when providing these services. Costs increase with increasing quantity provided. Your profit in Token is calculated by subtracting your costs from your payment.

A certain benefit for the patient is assigned to each quantity of medical services, the patient benefit that the patient gains from your provision of services (treatment). Therefore, your decision on the quantity of medical services not only determines your own profit, but also the patient benefit. An example for a decision situation is given on the following screen.

{Decision screen for patient 1C under FFS and CAP}

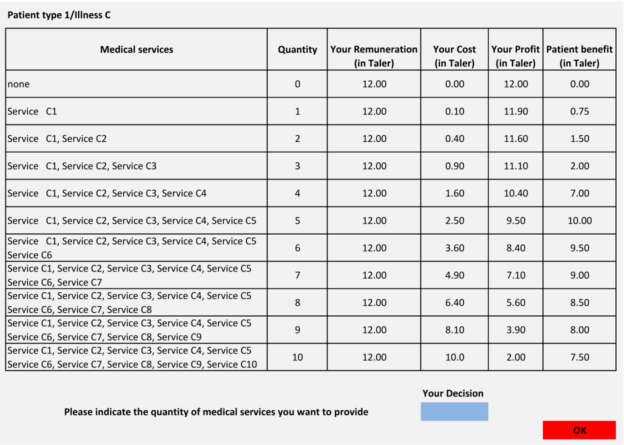


Patient type 1/Illness C

Medical services	Quantity	Your Remuneration (in Taler)	Your Cost (in Taler)	Your Profit (in Taler)	Patient benefit (in Taler)
none	0	0.00	0.00	0.00	0.00
Service C1	1	1.80	0.10	1.70	0.75
Service C1, Service C2	2	3.60	0.40	3.20	1.50
Service C1, Service C2, Service C3	3	5.40	0.90	4.50	2.00
Service C1, Service C2, Service C3, Service C4	4	7.20	1.60	5.60	7.00
Service C1, Service C2, Service C3, Service C4, Service C5	5	9.00	2.50	6.50	10.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6	6	10.80	3.60	7.20	9.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7	7	12.60	4.90	7.70	9.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8	8	14.40	6.40	8.00	8.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9	9	16.20	8.10	8.10	8.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9, Service C10	10	18.30	10.0	8.30	7.50

Please indicate the quantity of medical services you want to provide

Your Decision



Patient type 1/Illness C

Medical services	Quantity	Your Remuneration (in Taler)	Your Cost (in Taler)	Your Profit (in Taler)	Patient benefit (in Taler)
none	0	12.00	0.00	12.00	0.00
Service C1	1	12.00	0.10	11.90	0.75
Service C1, Service C2	2	12.00	0.40	11.60	1.50
Service C1, Service C2, Service C3	3	12.00	0.90	11.10	2.00
Service C1, Service C2, Service C3, Service C4	4	12.00	1.60	10.40	7.00
Service C1, Service C2, Service C3, Service C4, Service C5	5	12.00	2.50	9.50	10.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6	6	12.00	3.60	8.40	9.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7	7	12.00	4.90	7.10	9.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8	8	12.00	6.40	5.60	8.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9	9	12.00	8.10	3.90	8.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9, Service C10	10	12.00	10.0	2.00	7.50

Please indicate the quantity of medical services you want to provide

Your Decision

You decide on the quantity of medical services on your computer screen by typing an integer between 0 and 10 into the box labeled ”Your Decision”.

After all participants have taken their decisions for the respective patient you will proceed to the next patient.

Also in this part of the experiment there are no real, but abstract patients participating in this experiment. Yet, the patient benefit, which an abstract patient receives by your providing medical services, will be beneficial for a real patient. Also in the second part of the experiment the total amount of patient benefit determined by your 15 decisions will be provided to a patient with cancer treated in Shandong Qilu Hospital [Shandong University Cancer Hospital]. The money will be directly transferred to the patient’s in-hospital account to finance part of his/her treatment fee.

Each time you made a decision on the quantity of medical services you will be informed on your profit and the patient benefit. After you have made your 15 decisions in Part 2 of the experiment you will get to know your total profit and the corresponding total patient benefit.

Earnings in Part 2 of the experiment

After you have made your decisions in Part 2 of the experiment, your overall earnings will be calculated by summing up your profits from providing medical services to the 15 patients. This amount will be converted from Token into RMB at the end of the experiment and will be paid to you in cash together with the earnings of Part 1 of the experiment (rounded to 1 Yuan).

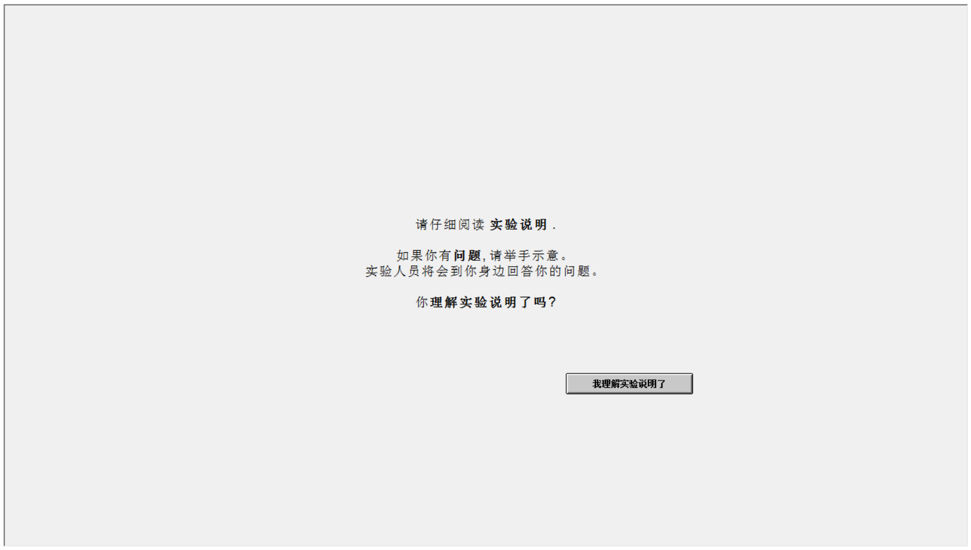
The patient benefit gained by all 15 patients will be converted into RMB at the end of the experiment, too, and will be transferred to the real patient’s in-hospital account. To this end the experimenter and a monitor will go together to Shandong Qilu Hospital [Shandong University University Hospital]. After the transfer, the signed receipt will be scanned into electronic form and will be sent to all the participants via e-mail in order to ensure the authenticity of the above process. Personal information will be blinded black to respect the patient's privacy. Information about the procedure has been given in Part 1 of the experiment.

Next, please answer some questions in this part of the experiment that will familiarize you with the present decision situation. After your 15 decisions, please answer some further questions on your screen.

C2: Comprehension questions prior to the experiment

The following example applies to FFS. For CAP, screens 3 to 5 are similar to Figure C1b in Appendix C4.

Screen 1



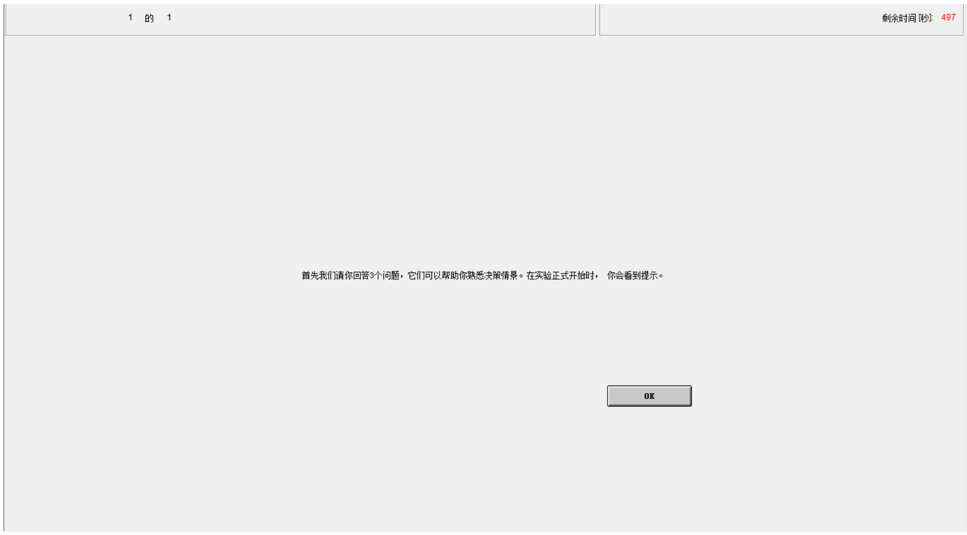
请仔细阅读 **实验说明** .

如果你有**问题**, 请举手示意。
实验人员将会到你身边回答你的问题。

你**理解实验说明了吗?**

Please read the instructions carefully. If you have a question, please raise your hand. The experimenter will come to you and answer your question. Have you understood the instructions?

Screen 2



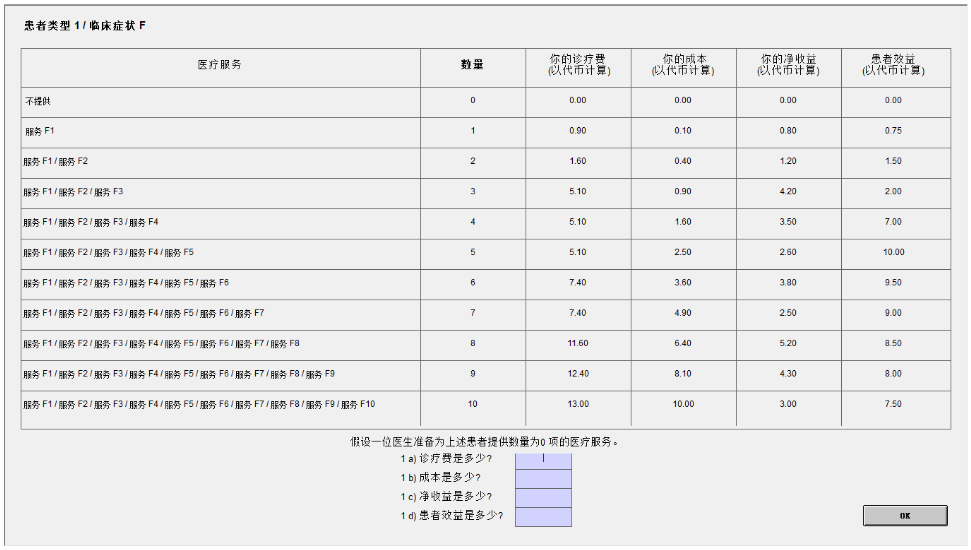
1 的 1

剩余时间 [秒]: 497

首先我们请你回答3个问题，它们可以帮助你熟悉决策情景。在实验正式开始时， 你会看到提示。

To familiarize you with the decision situation we first ask you to answer 3 questions. We will inform you when the actual experiment starts.

Screen 3 [4, 5]



患者类型 1 / 临床症状 F

医疗服务	数量	你的诊疗费 (以代币计算)	你的成本 (以代币计算)	你的净收益 (以代币计算)	患者效益 (以代币计算)
不提供	0	0.00	0.00	0.00	0.00
服务 F1	1	0.90	0.10	0.80	0.75
服务 F1 / 服务 F2	2	1.60	0.40	1.20	1.50
服务 F1 / 服务 F2 / 服务 F3	3	5.10	0.90	4.20	2.00
服务 F1 / 服务 F2 / 服务 F3 / 服务 F4	4	5.10	1.60	3.50	7.00
服务 F1 / 服务 F2 / 服务 F3 / 服务 F4 / 服务 F5	5	5.10	2.50	2.60	10.00
服务 F1 / 服务 F2 / 服务 F3 / 服务 F4 / 服务 F5 / 服务 F6	6	7.40	3.60	3.80	9.50
服务 F1 / 服务 F2 / 服务 F3 / 服务 F4 / 服务 F5 / 服务 F6 / 服务 F7	7	7.40	4.90	2.50	9.00
服务 F1 / 服务 F2 / 服务 F3 / 服务 F4 / 服务 F5 / 服务 F6 / 服务 F7 / 服务 F8	8	11.60	6.40	5.20	8.50
服务 F1 / 服务 F2 / 服务 F3 / 服务 F4 / 服务 F5 / 服务 F6 / 服务 F7 / 服务 F8 / 服务 F9	9	12.40	8.10	4.30	8.00
服务 F1 / 服务 F2 / 服务 F3 / 服务 F4 / 服务 F5 / 服务 F6 / 服务 F7 / 服务 F8 / 服务 F9 / 服务 F10	10	13.00	10.00	3.00	7.50

假设一位医生准备为上述患者提供数量为0 项的医疗服务。

1 a) 诊疗费是多少?

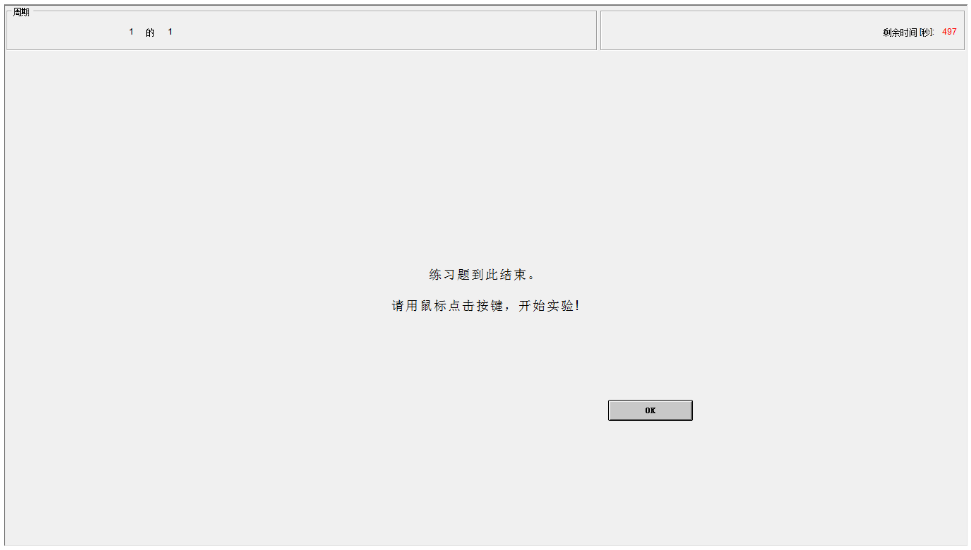
1 b) 成本是多少?

1 c) 净收益是多少?

1 d) 患者效益是多少?

Assume a physician wants to provide the quantity of 0 [10, 4] medical services for the patient above.
1 [2, 3] a) What is the remuneration?
1 [2, 3] b) What are the costs?
1 [2, 3] c) What is the profit?
1 [2, 3] d) What is the patient benefit?

Screen 6



周期

1 的 1

剩余时间 [秒]: 497

练习题到此结束。

请用鼠标点击按键，开始实验!

The test questions are now completed. When you click on the button the experiment will start!

C3: Questionnaires after Part 1 and Part 2 of the experiment.

[Information in brackets was requested from doctors only.]

{{Numbers/text in double braces refer to Part 2 of the experiment.}}

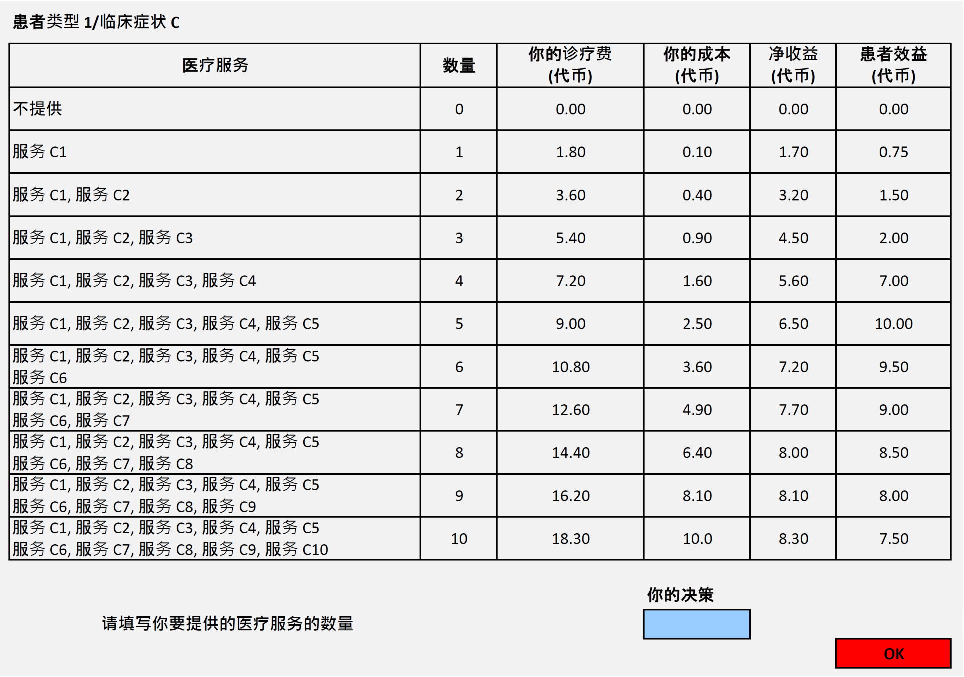
Please confirm your terminal number on your questionnaire. After you have made all decisions in Part 1 {{2}} of the experiment we would like to ask you to answer the following questions as good as possible. These answers are extremely important for our studies. Thank you for your cooperation.

Please put yourself back into the decision situation of Part 1 {{2}} of the experiment.

- What factors did influence your decision? Why did you decide in this way?
- How did the profit influence your decision?
- How did the patient benefit influence your decision?
- {{Major (faculty / main subject(s)):}}
- {{What is the number of your semester?}}
- {{Your gender: female/male}}
- {{Your nationality: (students only)}}
- {{[Your age]:}}
- {{[How many years of professional experience do you have?]}}
- {{[Your specification
- General Practitioner
- Traditional Chinese Medicine
- Public Health
- Other]}}

C4. Chinese decision screens.

Figure C1a. Illustration of the decision screen for patient 1C under CAP



患者类型 1/临床症状 C

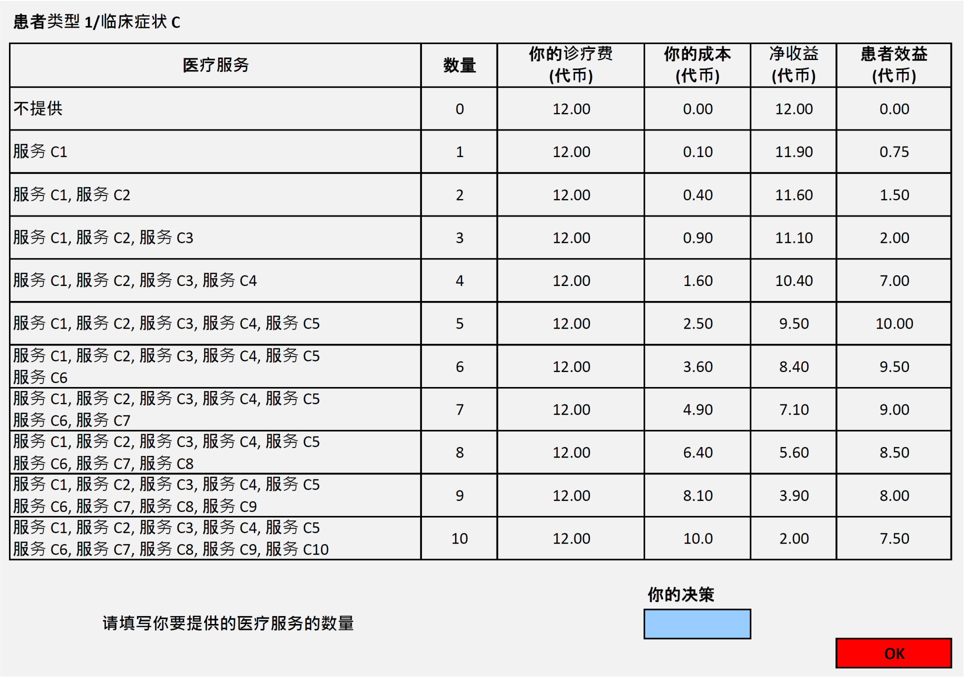
医疗服务	数量	你的诊疗费 (代币)	你的成本 (代币)	净收益 (代币)	患者效益 (代币)
不提供	0	0.00	0.00	0.00	0.00
服务 C1	1	1.80	0.10	1.70	0.75
服务 C1, 服务 C2	2	3.60	0.40	3.20	1.50
服务 C1, 服务 C2, 服务 C3	3	5.40	0.90	4.50	2.00
服务 C1, 服务 C2, 服务 C3, 服务 C4	4	7.20	1.60	5.60	7.00
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5	5	9.00	2.50	6.50	10.00
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6	6	10.80	3.60	7.20	9.50
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6, 服务 C7	7	12.60	4.90	7.70	9.00
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6, 服务 C7, 服务 C8	8	14.40	6.40	8.00	8.50
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6, 服务 C7, 服务 C8, 服务 C9	9	16.20	8.10	8.10	8.00
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6, 服务 C7, 服务 C8, 服务 C9, 服务 C10	10	18.30	10.0	8.30	7.50

你的决策

请填写你要提供的医疗服务的数量

OK

Figure C1b. Illustration of the decision screen for patient 1C under FFS



患者类型 1/临床症状 C

医疗服务	数量	你的诊疗费 (代币)	你的成本 (代币)	净收益 (代币)	患者效益 (代币)
不提供	0	12.00	0.00	12.00	0.00
服务 C1	1	12.00	0.10	11.90	0.75
服务 C1, 服务 C2	2	12.00	0.40	11.60	1.50
服务 C1, 服务 C2, 服务 C3	3	12.00	0.90	11.10	2.00
服务 C1, 服务 C2, 服务 C3, 服务 C4	4	12.00	1.60	10.40	7.00
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5	5	12.00	2.50	9.50	10.00
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6	6	12.00	3.60	8.40	9.50
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6, 服务 C7	7	12.00	4.90	7.10	9.00
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6, 服务 C7, 服务 C8	8	12.00	6.40	5.60	8.50
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6, 服务 C7, 服务 C8, 服务 C9	9	12.00	8.10	3.90	8.00
服务 C1, 服务 C2, 服务 C3, 服务 C4, 服务 C5 服务 C6, 服务 C7, 服务 C8, 服务 C9, 服务 C10	10	12.00	10.0	2.00	7.50

你的决策

请填写你要提供的医疗服务的数量

OK

D Results from mixed logit regressions

The experimental data can be analyzed by means of a standard mixed logit model. We have estimated a mixed logit model where the scale parameter is (silently, by default) constrained to 1 when $t = 1$. The scale parameter under $t = 2$ is identified as proportional to scale when $t = 1$. We describe and discuss the results below. We note that the mixed logit model does not provide results that lead to a different conclusion with regard to whether preferences are stable over subject pools. The conclusion with regard to how *experience* in the laboratory affects randomness in behavior is also unaffected by the specification of a mixed logit model. Estimation results are presented in Table D1. Subscripts “c” refer to Chinese students, “d” to Chinese doctors, “g” to German students.

Table D1: Results from maximum simulated likelihood

Sample: 178 Chinese students, 99 Chinese doctors, 42 German students. 30 (15) decisions for each Chinese (German) subject. Subjects are more experienced when $t = 2$. Preference parameters for patient benefit (α) and profit (γ) are assumed normally distributed with common standard deviation across the three subject pools.

	Chinese student	**Chinese doctor**	**German student**
MEAN			
α_n	8.12* CI(5.56 -10.67)	5.60* CI(4.00 - 7.99)	7.62* CI(4.41 - 10.83)
γ_n	3.72* CI(2.58 -4.87)	2.73* CI(1.78 - 3.69)	4.78* CI(2.96 - 6.62)
SD			
α_n	4.49 CI(3.17 -5.81)		
γ_n	2.47 CI(1.58 -3.35)		
$\frac{\sigma_1}{\sigma_2}$†	1.99		(N.A)

Note: Standard errors are clustered at the level of each individual when computing CI.

* Estimated parameter is significantly different from zero with a p-value < 0.001

† $\frac{\sigma_1}{\sigma_2}$ is significantly larger than 1 with a p-value < 0.001

In Table D1, σ_1 (σ_2) denote the unknown σ in the first (second) half of the experiment. In the model, we allow for $\sigma_1 \neq \sigma_2$. By allowing for randomness in behavior to depend on experience, and distribution of preference parameters to be unaffected by experience, we may identify the relative value of the σs from the two parts of the experiment. The interpretation of the results from our mixed logit model allowing for randomness in behavior to depend on experience is that $\sigma_1 > \sigma_2$ and hence, that subjects’ behaviors become less random as subjects get more experienced.

To examine differences in preferences by means of choice models, we examine the marginal rate of substitution (MRS), since this measure is independent of σ. The ratio of two

Table D2. Mean of |MRS| and their 95 % CI.

		95% CI	
Subject pool	|MRS|	L	H
Delta method:			
C students	2.180	1.737	2.623
Doctors	2.193	1.649	2.738
G students	1.592	0.963	2.220
Bootstrap†:			
C students	2.180	1.769	2.718
Doctors	2.193	1.708	2.897
G students	1.592	1.029	2.427

† Krinsky-Robb parametric bootstrap with 3000 replications.

coefficients is proportional to MRS when variables are log-transformed, and equal to the absolute value of MRS whenever patient benefit is equal to profit. Since the experiment is the same for all subject pools, we may compare the MRS for the three subject pools. Based on our regression estimates, we compute, for each subject pool, the means of coefficient ratios, $\frac{\beta_B}{\beta_\pi}$, and their 95 % confidence intervals, using the STATA module WTP by Arne Risa Hole. Confidence intervals are computed in alternative ways, applying both the Delta method and the Krinsky-Robb parametric bootstrap method. Table D2 reports the estimated absolute values of means of marginal rates of substitution (in the case where patient benefit is equal to profit), and the corresponding 95 % confidence intervals. Comparing MRS over the subject pools, we find that we cannot reject the hypothesis that the MRSs are equal for all three subject pools.