Production, Performance, and Payment:

Five Essays on Incentive Contracts

Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften

(Dr.rer.pol.)

durch die Fakultät für Wirtschaftswissenschaften der

Universität Duisburg-Essen

Campus Essen

Vorgelegt von

Name: Nadja Vanessa Kairies-Schwarz

Geburtsort: Sinsheim

Essen 2013

Tag der mündlichen Prüfung: 29.08.2013

Erstgutachterin: Prof. Dr. Jeannette Brosig-Koch

Zweitgutachter: Prof. Dr. Harald Tauchmann

Contents

Introduction

Introduction

INCENTIVES

"*People are people, and they respond to incentives. They can nearly always be manipulated -for good or ill - if only you find the right levers*." (Levitt and Dubner, 2009). Incentives play a central role in our lives. Even if we do not always notice them immediately, we are all guided by them in everyday life. For example, we are offered coupons for a certain store which encourage us to buy our groceries there, we donate blood because afterwards we feel good about ourselves, and we work long hours at a consultancy to meet the ideal of a corporate culture.

These examples indicate that the types and mechanisms of incentives are manifold and that they work by reference to different underlying human motivations. According to Bénabou and Tirole (2006), human motivations can be categorized as intrinsic, extrinsic, and reputational. While individuals are intrinsically motivated by a task itself, e.g. in the case of altruism, extrinsic motivation can be derived from a monetary or material reward.[1] Reputational motivation is based on the assumption that people are motivated by signaling behavior that is perceived as good by society, thus conforming to social pressure or norms. Good behavior can either be signaled to oneself in order to uphold a certain self-image and self-esteem (Bénabou and Tirole, 2002, Dana et al., 2007), or to others in order to be perceived as good by others and receive a positive social reputation (Bénabou and Tirole, 2006, Rege and Telle, 2004, Ariely et al., 2009). In practice, extrinsic and reputational incentives are frequently implemented in order to change a person's behavior, and all of the incentives above mentioned imply either monetary rewards or reputational concerns. Coupons are an extrinsic monetary incentive, donating blood involves reputational motivation in the sense of signaling good behavior to oneself, and working long hours at a consultancy brings reputational motivation in the sense signaling good behavior to others.

Economists use incentives widely to change the behavior of economic agents towards more desirable outcomes. In the last decades, incentives have played an important

[1] Regarding the interplay between these two motivations, both theoretical models (Bénabou and Tirole, 2003) and experimental evidence (see e.g., Frey and Oberholzer-Gee, Gneezy and Rutichini, 2000) show that extrinsic incentives can crowd out intrinsic motivation.

role especially for labor relationships, where the interests of the firm and of the employee are often not aligned. In order to motivate people to act in their interest, firms typically embed incentives within compensation schemes. In recent years, incentives that reward performance have become especially popular in business practice. Just recently, the new Yahoo chief executive Marissa Mayer was awarded a highly performance-dependent payment package. It includes a base salary of $1 million and various performance-based payments such as a bonus of up to $4 million if predetermined performance measures are met.[2]

When individual performance is easily observable, i.e. when the executed task is simple, empirical evidence finds that performance incentives matter. Lazear (2000) shows that output increased by 44 % when compensation for workers of a glass manufacturer was changed from salary to piece rate. Paarsch and Shearer (2000) indicate that productivity in a tree-planting firm was higher by 22.6% for workers who received piece rate compensation than for those who received a fixed payment. Bandiera et al. (2007) show that productivity increased by 21% when the compensation of managers for a soft fruit producer in the UK was changed from a fixed wage to piece rate.

However, for most workers performance is not easily measureable due the complexity of the task to be executed (Pendergast, 1999).

RESEARCH QUESTIONS

The aim of this dissertation is to empirically and theoretically investigate the effects of different performance incentives on individual distribution and provision behavior in such complex production situations. Chapter 1, Fischbacher et al. (2013), analyzes whether the salience of performance measures in complex team production situations plays an important role for perceptions of distributive fairness by means of a laboratory experiment. Chapters 2-5, i.e. Brosig-Koch et al. (2013a, b), Kairies and Krieger (2013), and Kairies (2012), investigate whether and how different monetary and non-monetary incentives affect provision behavior of professionals whose

[2] See New York Times: http://dealbook.nytimes.com/2012/07/27/adding-up-marissa-mayers-pay-at-yahoo/ (retrieved:04.02.2013)

performance is multidimensional and difficult to measure – in this case physicians – by using laboratory experiments and a theoretical model.

TEAM PRODUCTION: WHO CONTRIBUTES?

First, complexity of a job may arise when many workers execute tasks for which the outcome is not solely the result of individual inputs, but also of joint inputs by multiple workers. Examples of joint production where individual contributions are not easily measurable include joint ventures, government agencies, law firms, teaching, and medical practice. As individual contributions in work teams are often difficult to observe, employees are frequently paid a fixed share of the total outcome. However, compensation schemes which do not account for individual performance may lead to shirking and inefficient effort levels due to free-riding (Alchian and Demsetz, 1972, Holmstrom, 1982).

When performance is not fully observable firms sometimes make use of subjective performance measures (Rajan and Reichelstein, 2006). Subjective performance measures, however, often lead to rent-seeking activities by supervisors (Holmstrom 1982, Milgrom and Roberts, 1988), which in turn may cause disputes among employees and even, e.g., costly strikes. In order to forgo such disputes and design adequate subjective performance measures for team production, it is hence important to understand people's perceptions of distributive fairness in joint production situations.

MULTIPLE ASPECTS OF PERFORMANCE: WHAT CAN BE OBSERVED?

Second, performance is often difficult to measure when tasks or jobs are more complex in the sense that they have more than one outcome dimension, e.g. not only quantity but also quality. For hairdressers, for instance, it might not only be important how many haircuts they give per day, but also how nice and well cut they are. When only few dimensions are actually measureable and included in performance metrics, employees may substitute incentivized aspects for non-incentivized aspects as a result of multitasking (Holmstrom and Milgrom, 1991, Baker 1992, Courty and Marschke, 2004). Typical professional groups in which it is difficult to measure all dimensions of performance include lawyers, teachers, and physicians.

Especially the compensation of the latter has drawn much media attention in Germany in recent months. In the city of Goettingen, e.g., a physician at a hospital forged patient data in order to receive donor organs, as performance incentives were based on the quantity of organ transplants conducted.[3] Performance payment dependent on the quantity of medical services is common in health care, as outcomes in terms of patient well-being are difficult to observe. Typically, physicians have been remunerated by compensation schemes such as fee-for-service or capitation. Under fee-for-service, a physician receives a fee for each quantity of medical treatment provided, and under capitation he receives a lump-sum payment per capita irrespective of the quantity of medical treatment provided. From a theoretical point of view these incentives are thus contrary to each other. While a fee-for service compensation leads to the overprovision of medical treatment, capitation leads to underprovision (Ellis and McGuire, 1986). Payment reforms targeting under- and overprovision include mixed payment systems which are usually a combination of these basic payment schemes (Krasnik et al., 1990, Iversen and Lurås, 2000, Dumont et al., 2008).

Neither the classic payment schemes nor mixed models, however, usually target what ought to be most important: the patient outcome. As President Barack Obama of the United States of America put it in his speech to the Centers of Medicare and Medicaid in 2009: "*You did not enter this profession to be bean counters and paper pushers. You entered this profession to be healers*".[4] Relating this notion to physician payment suggests that future compensation schemes should take this most fundamental aspect of the medical profession, the quality of care, into account. Reforms targeted at improving the quality of care include pay-for-performance incentives, or paying physicians to fulfill certain predetermined performance measures of which some are outcome related (Doran et al., 2006, Rosenthal, 2008). Non-monetary incentives are even more frequently used in attempts to improve the quality of care. Non-monetary performance incentives are based on the theoretical assumption that people are not solely motivated by money but also care about their self-esteem (Bénabou and Tirole,

[3] See e.g. Frankfurter Allgemeine Zeitung: http://www.faz.net/-gqe-721ce (retrieved: 12.08.2012)
[4] See e.g. in the New York Times: http://www.nytimes.com/2009/06/16/health/policy/16obama.html?ref=politics&pagewanted=print&_r=0 (retrieved: 21.01.2012)

2002), i.e. they like being a good physician, or their social reputation (Bénabou and Tirole, 2006), i.e. they like being *perceived* as being a good physician by others. In practice, policy makers have started to implement private and public feedback systems in several countries. While private feedback speaks to self-esteem, public feedback also addresses the desire to be perceived as a good physician.[5]

THE METHOD OF ECONOMIC EXPERIMENTS

Using experiments in the economic sciences is a rather new development. In contrast to other sciences such as biology, physics or chemistry, economic science was purely observational and theoretical for a long time (Cronson and Gächter, 2010). While economic theories offer a good framework to analyze and describe economic situations, they are limited as they are an abstraction and are based on behavioral assumptions. Generally speaking, behavioral assumptions include economic agents' evaluation of outcomes implied in their preferences, their cognitive abilities such as decisions under uncertainty, and their behavior usually described in equilibrium solution concepts. Despite the fact that economic theories are only an abstraction, they are useful as they provide testable and thus falsifiable (see e.g., Popper 1934, 1963) predictions. To test the predictions of theories one needs to collect data. However, natural experiments which guarantee control of the environment are very rare. Control means that only the parameters of interest (the treatments) are varied and hence changes in behavior can be attributed only to the changes in these parameters. A further problem is that in order to truly test a theory, an experiment must test not only the predictions of the model but also its assumptions – a predicament known as the Duhem-Quine problem. In the case that the predictions of an experiment are not verified, one cannot say whether this is due to a violation of the assumptions made, or to a wrong prediction of the theory. In this case, laboratory experiments have an advantage over using field data, as assumptions and predictions can be tested separately.[6]

[5] In the long-term, public feedback can also have financial consequences for physicians.
[6] However, laboratory experiments often cannot entirely avoid this problem, as they rely on certain assumptions as well, e.g. that the subjects' utilities depend primarily on their payoff.

According to Smith (2002), the ideal method for conducting laboratory experiments is to first derive a testable prediction from a well-specified theory and then implement an experimental design that implicitly includes auxiliary assumptions which have been tested in previous experiments. The latter include using relevant and clear instructions about what the subjects in an experiment can do, what they know, how their decisions are transformed into monetary payoffs guaranteeing that subjects take their decisions seriously, and randomly assigning the subject pool, usually consisting of students. In short, laboratory experiments are an abstraction of real world economic decisions, just like theories. Compared to data collection in the field, however, they have the advantages of control over parameters and auxiliary assumptions and a less biased subject pool. Following Roth (1995), economic experiments are conducted for roughly three purposes: to test theories ("Speaking to Theorists"), to observe new regularities of human behavior ("Searching for Facts"), and as test beds for policy instruments ("Whispering in the Ears of Princes").

The aim of experiments of the first type is to analyze existing theories and give feedback, hopefully engendering a fruitful dialogue between experimenters and theorists. Economists started to conduct formal economic experiments to test theories in the first half of the 20th century. Early experiments included the analysis of individual choice behavior, such as testing the indifference curve (Thurstone, 1931). Another substantial impulse for experimental economics was given by von Neumann and Morgenstern's "Theory of Games and Economic Behavior" (1944). Their expected utility theory provided a basis for new experiments which tested game theoretic predictions (Mosteller and Nogee, 1951, Allais, 1953, Flood, 1952). In a third experimental wave in the 1960s economists then began testing theories of industrial organization, e.g. on bargaining in oligopolies (Sauermann and Selten, 1960), duopolies, and bilateral monopolies (Siegel and Fouraker, 1960).

The second category of economic experiments evolved through ongoing experimentation. Experimenters often observe robust regularities of behavior which have not yet been explained by theory. Ideally, these observations then lead to the development of new theories. Observations from experiments led, e.g., to the development of prospect theory (Kahnemann and Tversky, 1979), or theories of inequity aversion (Fehr and Schmidt, 1999, Bolton and Ockenfels, 2000). Theories of inequity aversion are a prime example, as they themselves provided new predictions which were tested (Engelmann and Strobel, 2004). As a consequence of observed inconsistencies, new theories of other-regarding preferences were developed (Charness and Rabin, 2002, Falk and Fischbacher, 2006). Thus, experimentation leads to a continuous dialogue between theorists and experimenters with the aim of improving theories.

The latest category of the three comprehends experiments which provide a test bed for policy makers. Such experiments are usually motivated by the intention of decision makers to change institutions. Beyond control and random subject pools, laboratory experiments also offer the advantage of being a relatively cost-efficient means of testing new institutions. Policy changes usually concern market designs, including for instance a kidney exchange system (Roth et al. 2004, 2005), school and university admissions (Abdulkadiroglu et al., 2005, Dwenger et al., 2010), spectrum auctions (Grimm et al., 2003), or online auctions (Bolton et al., 2004, Ariely et al., 2005).

According to Roth (1988) economic experimentation has started its "*steady and sustained transformation from an occasional curiosity into a regular means for investigating many kinds of economic phenomena*". The success of this method can be underlined by the fact that the share of experimental papers published in the top three economics journals has risen from 0.84% -1.58% in the 1980s to 3.80-4.15% between 2000 and 2008 (Falk and Heckmann, 2009).[7] However, experiments still face resistance. One point of criticism is that laboratory experiments lack realism and thus many economists argue that field experiments are of greater relevance to policy (Levitt and List, 2007, 2009). In particular, they criticize that in laboratory experiments subjects are scrutinized, there is a deficiency of anonymity between experimenter and subject, decision situations often lack context dependency, stakes are not realistic, and only certain types of subjects select into experiments. However, Falk and Heckman (2009) argue that realism is not the real issue in the debate over "laboratory vs. field", but rather the possibility of isolating causal effects – which in turn calls for more laboratory experimentation. Examples for laboratory experiments trying to isolate

[7] The three top journals include The American Economic Review, Econometrica, and The Quarterly Journal of Economics.

causal effects include research analyzing games with high stakes, or using subjects who are not students and, e.g., CEOs (Fehr and List, 2004).

The method of experimental economics is thus ideal for analyzing complex production situations. While theoretical models on performance incentives in the workplace are abundant, testing them in the field is often difficult when production processes or job tasks are more complex. This holds especially in team production and for tasks for which not all dimensions of performance are measurable and environmental influences are difficult to control for. In the context of team production where individual contributions are often difficult to observe, we use a laboratory experiment to control for the complexity of the production process. In the field of health economics laboratory experiments are relatively new and few studies exist so far (an early example is Hennig-Schmidt et al., 2011). Using the method of laboratory experiments to analyze physician compensation schemes has several advantages. In a laboratory experiment we can investigate individual decision behavior before and after a payment reform by varying the performance incentives and keeping everything else constant. In addition, we are able to assume that quality is fully observable and hence exclude the problem of multitasking. To control for external validity, we also conduct laboratory experiments with medical students.[8] Experimental economics may thus offer health policy makers a cost efficient way to analyze the effects of planned reforms before they are implemented and thus help to avoid costly policy failures.

In the following I summarize Chapters 1-5 of my dissertation.

CHAPTER 1

In Chapter 1, we address the problem of performance measurement in complex team production. In team production processes, mutually agreed-upon distribution rules for the ex post allocation of the joint product are vital to providing team members with adequate monetary incentives ex ante. However, if team members regard the distribution of income as unfair their motivation might be negatively affected. The degree to which individual performance is observable determines the availability of performance measures as a basis for monetary incentives, but it also affects team

[8] We also began conducting laboratory experiments with actual physicians, though not enough observations are available yet to perform meaningful analyses.

member's norms of distributive fairness. The degree to which performance measures are obvious occurs especially in complex team production situations in which individual contributions to the joint output are often difficult to observe, for instance in work teams, co-authored publications, joint ventures, or mergers or acquisitions. For the latter, Cronson et al. (2004) for example find that they may result in both synergies and negative externalities, thus making it difficult to observe individual contributions. Recent experimental evidence shows that team members allocate their effort according to distributive norms which account for individual contributions (Konow, 2000, Gantner et al., 2001, Cappelen et al., 2007, Cappelen, et al., 2008, Konow et al., 2009). The most prominent distribution norms include equal sharing based on the fairness ideal of strict egalitarianism (Nielsen, 1985), accountability in the sense of holding people responsible only for factors in their control (Konow, 1996, 2000), and marginal productivities based on the libertarian fairness ideal (Nozick, 1974). However, only few studies can accommodate the complexity of a production process. While Gächter and Riedl (2005) account for negative externalities and Karagözoğlu and Riedl (2011) for uncertainty in the production process – they show that entitlements play a role under these circumstances – they are not able to investigate the prevalence of distributive norms.

However, as it is crucial from a motivational point of view to anticipate perceptions of distributive fairness for the design of performance measures, we use a laboratory experiment to analyze what people consider a fair distribution when individual contributions are not readily apparent. The experimental design is as follows: Joint production for a team of two takes place by a neutral third party. Team members can increase joint output by answering a knowledge question. Thus, effort costs are not observable. Varying the increase in the joint output resulting from each team member giving the correct answer allows us to investigate whether individual inputs (i.e. giving the right answer per se) or marginal productivities (i.e. the increase in points realized by giving the correct answer) affect the distribution chosen. In each of the nine production situations we showed our participants a payoff matrix in which three situations were additive, three allowed for synergies, and three for negative externalities. Hence, individual marginal productivities were noticeable, but not always apparent.

Our results indicate that people hold others responsible for their inputs, i.e. they distribute according to the accountability principle. In particular in the asymmetric cases, third parties allocate more points to the subject who knew the correct answer to the knowledge question. In the symmetric cases, in which both subjects either know or do not know the correct answer, they distribute equally. The predominance of the accountability principle is robust to the role of the person distributing the joint production payoff, i.e. whether it is the neutral third party or the proposers and responders in an Ultimatum Game, and to the framing of the production process, i.e. whether it is simultaneous or sequential. Although both our framings are economically equivalent, marginal productivities are only considered if they are made apparent. Hence, the results offer an explanation for why subjective performance measures are frequently regarded as unfair.

CHAPTER 2

In chapters 2-5, performance incentives for the medical profession are analyzed. As mentioned earlier, the treatment quality provided by physicians is difficult to measure and thus difficult to control. Abstracting from measurement difficulties, the effects of different monetary and non-monetary performance incentives for physicians are analyzed by means of laboratory experiments in Chapters 2-4 and by a theoretical model in Chapter 5.

In Chapter 2, we derive an experimental design of performance incentives for physicians and analyze provision behavior under the basic incentive schemes of capitation and fee-for-service as well as various mixtures of the two.

Concerning the basic incentives, theory suggests the effects of underprovison given capitation and overprovision given fee-for-service are mitigated in mixed remuneration systems (Ellis and McGuire, 1986). Hennig-Schmidt et al. (2011) provide empirical evidence which supports this. However, their experimental design involves a rather complicated parameterization of payment incentives and patient health benefit. As a consequence, it lacks comparability across payment incentives and different patient types. Moreover, they use an across-subject design which does not allow for within-subject analysis. Our study models a much more general framework for the classic payment schemes of capitation and fee-for-service, with patient benefits that allow for a clear characterization of illness and severity of illness.

In order to mitigate the effects of under- and overprovision, several countries have introduced mixed payment systems. Empirical evidence confirm that mixed payment incentives can have this effect, see e.g. Krasnik et al. (1990) for Denmark, Iversen und Lurås (2000) for Norway or Dumont et al. (2008) for Canada. These field studies, however, lack control over some aspects such as the variation of multiple parameters of the payment systems, self-selection of physicians into new payment systems, and missing acknowledgement of institutional characteristics or the health status of patients treated. Hence, they are not controlled analyses of the effect of mixed payment incentives on physician provision behavior.

To complement the existing field studies, we conduct a controlled laboratory experiment. In this experiment subjects take on the role of a physician and decide on the quantity of medical treatment for nine different patient types which result from the combination of three different illnesses (*A*,*B,C*) and three different severities of these illnesses (*x*,*y*,*z*). Patients are not present in the laboratory. However, subjects are informed that the monetary value of the patient benefit arising from their treatment decision goes to real patients who are treated for eye cataract. Thus, their decisions enable surgery for real patients. To account for the character of a payment reform we use a within-subject design. In part 1 of the laboratory experiment, subjects decide on the quantity of medical treatment for the nine patients under either capitation or fee-for-service remuneration. In part 2, they are then faced with different mixtures of these two basic payment systems, which diverge from the basic remuneration scheme in part 1 to varying degrees.

We find that mixed payment incentives significantly dampen the effects of over- and underprovision observed under fee-for-service and capitation. Hence, patients are better off under mixed payment than under the basic payment systems. However, considering efficiency, i.e. cost-benefit analysis, we find no significant improvement. Hence, the increase in money put into the system cannot be outweighed by the increased patient benefits. Concerning the patient types, we find that severities play a significant role, whereas illnesses do not.

CHAPTER 3

In Chapter 3 we analyze whether pay-for-performance (P4P) incentives improve patient care compared to the pure payment systems of capitation and fee-for service. P4P incentives explicitly target the quality of care and not simply the quantity of treatment provided. They have been implemented in several countries such as the UK (Doran et al., 2006, Campbell et al., 2009), the USA (Rosenthal et al., 2004, and Rosenthal, 2008), Australia (Duckett et al. 2008, Scott, 2008, Scott et al., 2009), Korea (Kim, 2012), New Zealand (Perkins and Seddon, 2006, and Buetow, 2008), and Spain (Gené-Badia et al., 2007). Despite the popularity of P4P incentives, good data for evaluation is limited and it is yet an unanswered question, whether P4P incentives actually lead to an increase in the quality of care (Maynard, 2012). Existing evidence on early implementations generally shows that the effects on the overall quality of medical care were rather moderate; see e.g. Mullen et al. (2010) for the California P4P scheme or Campbell et al. (2009) and Gravelle et al. (2010) for the Quality and Outcomes Framework in the UK. A possible reason for the ambiguous effects may be negative side-effects due to design problems arising from the complexity of measuring quality. These include substitution away from non-incentivized to incentivized quality aspects of care (Mullen et al., 2010), improving quality to the performance target (Campbell et al., 2009), or gaming quality indicators by exception reporting (Gravelle et al., 2010). This raises the question whether the implemented quality metrics were not reliable, or whether perhaps P4P incentives fail to work at all.

Our experiment serves as a first step towards understanding the effects of P4P incentives within a medical context in a controlled environment. We build on the experimental design on the basis of capitation and fee-for-service remuneration outlined in Chapter 2. The design is again sequential, i.e. subjects in the role of a physician decide on the quantity of medical treatment for nine different patients, under a basic payment scheme in the first part of the experiment and under pay-for-performance payment in the second part. We attempt to capture the predominant characteristics of pay-for-performance incentives in the field in our design of the incentive mechanism. The performance incentive is thus modeled as an additional bonus given on top of the existing basic payment. Physicians can receive a bonus when they provide good performance. The performance measure is based on treatment quality, i.e. patient outcomes, which is perfectly observable in our experiment. The bonus is paid when the physician provides a quantity of medical treatment within a predefined interval around the around the patient-optimal level. We conduct our experiments with non-medical and medical students as subjects and can therefore control for any differences in behavior which may result from professional norms.

We find that physician provision behavior under pay-for-performance incentives differs significantly from that under basic capitation and fee-for-service remuneration. Given additional performance incentives, physicians treat patients at levels much closer to their optimum, hence improving the quality of care. We also consider the benefit/cost ratio for P4P as compared to the basic payment schemes and find that the increase in the additional costs cannot be overcompensated by the increase in patient benefits. This result supports the findings of previous field studies (see e.g. Mullen et al. 2010). Provision behavior is robust for medical and non-medical students. Moreover, an analysis of individual behavior reveals that given financial P4P incentives the majority of physicians can be classified as profit maximizers. Thus, this supports or finding that physicians respond to financial P4P incentives by increasing the quality of care. Hence, our results lead us to conclude that financial P4P incentives are able to increase patient benefits, but do not improve cost-effectiveness.

CHAPTER 4

In Chapter 4, we experimentally investigate whether and how non-financial performance incentives affect physician provision behavior. Non-monetary incentives have become another interesting approach for policy makers, as they might lead to improvements in patient care at a much lower cost than monetary incentives (Dranove and Jin, 2010). Non-monetary performance incentives are typically introduced in the form of public quality reporting, and evidence shows that they lead physicians to change their provision behavior. Although some evidence suggests an improvement in quality for the reported aspects (Hannan et al. 1994, Rosenthal et al., 1997) public reporting also leads to unintended problems, leaving the effect on the overall quality of care ambiguous (Marshall, 2000). Such problems include substitution away from unreported aspects of care (Werner et al., 2009) or patient selection (Dranove et al., 2003, Cutler et al., 2004, Werner and Asch, 2005). Also, this type of incentive is sensitive to regional heterogeneity in the degree of market competition (Grabowski and Town, 2011) and simultaneous introduction with monetary incentives (Maynard, 2012). Consequently, Cutler et al. (2004) postulate the need for more research on non-monetary performance incentives in order to better understand the underlying mechanisms that drive changes in physician provision behavior.

We therefore investigate how non-monetary performance incentives for physicians affect the quality of medical services provided in a controlled laboratory experiment. This design, too, builds on the framework described in Chapter 2. Physicians decide on the quantity of medical treatment for nine different patients. In part 1 of the experiment subjects decide under fee-for-service incentives, while in part 2 they are informed that in addition to this they will receive either private or public feedback on their performance at the end of the experiment. We distinguish between private and public performance feedback as it is necessary from a theoretical point of view. While private feedback provides competitive incentives addressing self-esteem (Bénabou and Tirole, 2002), making relative performance known to others adds a reputational motivation (Bénabou and Tirole, 2006). We measure performance in terms of outcome quality for all patients and give feedback in the form of competitive rankings. At the end of the experiment subjects receive private feedback by being shown their individual position in the ranking of all attendees on their computer screen. For public feedback subjects stand up at the end of the experiment while the ranking is read out loud by the experimenter; this procedure is similar to those used by Rege and Telle (2004) and Ariely et al. (2009). We conduct this experiment with non-medical and medical students.

Our results show that public feedback incentives have an effect on the quality of medical care. We find that this effect depends on the feedback mode. While private feedback has no impact on the quality of care, public feedback has a significant positive impact. Our results are robust to a subject's degree program (medical studies or other), changes in the task order, and other socio-demographic factors.

CHAPTER 5

In Chapter 5, I construct a theoretical framework that provides insights into how monetary and non-monetary incentives work when they are implemented jointly.

In the last decades, especially monetary-incentives have been used to improve the quality of medical treatment in terms of patient outcomes. However, if physician payment includes high monetary incentives dependent on the quantity of treatment, overprovision is likely (Ellis and McGuire, 1986). This is the case e.g. within Medicare in the US, where physicians receive high fee-for-service payments. Overprovision is a crucial problem as it is not only costly, but may even harm patients' health, e.g. when excess x-rays or medications are subscribed. In a situation of overprovision, a policy maker's two prevalent goals of improving the quality of care while reducing its costs need not be mutually exclusive. In recent years health policy makers have started to make use of non-monetary incentives based on reputational motivation (Bénabou and Tirole, 2006). In contrast to monetary incentives, non-monetary incentives may have the advantage that they come at relatively low costs and may be an important means to improving treatment quality.[9] Non-financial incentives addressing social reputational concerns are predominantly implemented in the form of public quality reporting. Early evidence shows that physicians respond to these incentives (Dranove et al., 2003, Cutler et al., 2004, Werner and Asch, 2005).

In order to analyze monetary and non-monetary incentives and account for the scenario of overprovision of medical treatment, I extend the seminal paper by Siciliani (2009). In contrast to Siciliani, I focus the analysis on non-monetary incentives addressing reputational motivation. Additionally, I conduct an efficiency analysis and derive the efficiency maximizing price. The assumptions of the model can be summarized as follows. Monetary incentives are given by a fee-for-service for each quantity of medical treatment. When abstracting from reputational motivation, a higher fee-for-service leads to an increase in medical treatment as it increases physicians' marginal revenues. To account for reputational motivation, I introduce a patient benefit function which links the quantity of care to the corresponding patient benefit. I assume that physicians are altruistic, as in Ellis and McGuire (1986), and that they care about their reputation, as in Bénabou and Tirole (2006). Like Ellis and McGuire (1986), but in contrast to Siciliani (2009), I assume that patients are characterized by a peaked patient benefit function allowing for under- and overprovision of care. If physicians treat patients close to their benefit maximum, i.e.

[9] When neglecting the implementation costs of public reporting systems.

over- or underprovision are low, they receive an extra gain in utility by being perceived as “good”. Physicians are heterogeneous in their degree of altruism. Those who are more altruistic adjust the quantity of care in order to be perceived as “good”, while less altruistic physicians maximize their profits.

I find that in contrast to monetary incentives, better reputational motivation unambiguously reduces the magnitude of overprovision, which in turn increases patient benefit and decreases the costs of healthcare provision. Moreover, regarding the efficiency of the monetary incentive scheme, which trades-off physicians' demand for a high fee-for-service (higher marginal revenues) against the patients' interest in a low fee-for-service (less overprovision), I show that an efficiency-maximizing price exists and that it decreases in reputational motivation.

Literature

Abdulkadiroglu, A., Parag, Pathak, A., and A.E. Roth. 2005. ”The New York City High School Match.” *American Economic Review, Papers and Proceedings* 95, 364-367.

Alchian, A. and H. Demsetz. 1972. “Production, information costs, and economic organization.” *American Economic Review* 62, 777-795.

Allais, M. 1953. “Le Comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’é´cole américaine.” *Econometrica*, 21, 503-46.

Ariely, D., Ockenfels, A., and A. E. Roth. 2005. “An Experimental Analysis of Ending Rules in Internet Auctions.” *RAND Journal of Economics* 36(4), 890-907.

Ariely, D., Bracha, A., and S. Meier. 2009. “Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially.” *American Economic Review* 99, 544–55.

Bandiera, O., Barankay, I., and I. Rasul. 2007. “Incentives for managers and inequality among workers: evidence from a firm-level experiment.” *The Quarterly Journal of Economics* 122(2), 729-773

Baker, G. 1992. “Incentive Contracts and Performance Measurement.” *Journal of Political Economy* 100, pp. 598-614.

Bénabou, R. and J. Tirole. 2002. “Self-confidence and personal motivation.” *Quarterly Journal of Economics* 117(3), 871–915.

Bénabou, R. and J. Tirole. 2003. “Intrinsic and extrinsic motivation”.*The Review of Economic Studies* 70(3), 489-520.

Bénabou, R. and J. Tirole. 2006. “Incentives and Prosocial behavior.” *American Economic Review* 96, 1652-1678.

Bolton, G.E., Katok, E., and A. Ockenfels. 2004. “How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation.” *Management Science* 50(11), 1587-1602.

Braun, S. Dwenger, N., and D. Kübler. 2010. “Telling the Truth May Not Pay Off: An Empirical Study of Centralized University Admissions in Germany.” *The B.E. Journal of Economic Analysis & Policy* 10(1), 22.

Brosig-Koch, J., Hennig-Schmidt, H. Kairies, N., and D. Wiesen. 2013a. “How to Improve Patient Care? An Analysis of Capitation, Fee-for-Service, and Mixed Incentive Schemes for Physicians.” *Ruhr Economic Papers* No. 412.

Brosig-Koch, J., Hennig-Schmidt, H. Kairies, N., and D. Wiesen. 2013b. "How effective are pay-for-performance incentives for physicians? A Laboratory Experiment" *Ruhr Economic Papers* No. 413.

Buetow, S. 2008. "Pay-for-Performance in New Zealand Primary Health Care." *Journal of Health Organisation and Management* 22, 36-47.

Camerer, C.F. and R.M. Hogarth. 1999. "The Effect of Financial Incentives." *Journal of Risk and Uncertainty* 19(1- 3), 7-42.

Cappelen, Alexander, Erik Sørensen, and Bertil Tungodden. 2012. "Just Luck: An Experimental Study of Risk Taking and Fairness," forthcoming in *American Economic Review*.

Croson, R. and S. Gächter. 2010. "The science of experimental economics." *Journal of Economic Behavior & Organization* 73, 122-131.

Cutler, D.M., Ilckman, R.S., and M. B. Landrum. 2004. "The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery." *American Economic Review*, 94(2), 342-346.

Campbell, S.M, Reeves, D. Kontopantelis, E. Sibbald, B., and M. Roland. 2009. "Effects of Pay for Performance on the Quality of Primary Care in England." *New England Journal of Medicine* 361, 368-78.

Cappelen, A.W., Hole, A.D. Sørensen, E.Ø., and B. Tungodden. 2007. "The pluralism of fairness ideals: an experimental approach." *American Economic Review* 97(3), 818–827.

Cappelen, A.W., Moene, K.O. Sørensen, E.Ø., and B.Tungodden. 2008. "Rich meets Poor - An International Fairness Experiment." Tinbergen Institute Discussion Papers 08-098/3.

Cappelen, A.W., Sørensen, E.Ø., and B. Tungodden. 2010. "Responsibility for what? Fairness and individual responsibility." *European Economic Review* 54(3), 429–441.

Croson, R.T.A., Gomes, A., McGinn, K.L., and M. Nöth. 2004. "Mergers and acquisitions: an experimental analysis of synergies, externalities and dynamics." *Review of Finance* 8, 481–514.

Dana, J., Weber, R. A., and J. X. Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory* 33(1), 67-80.

Deci, E. L. and R. M. Ryan. 1985. “Intrinsic motivation and self-determination in human behavior. “ New York: Plenum.

Doran, T., Fullwood, C. Gravelle, H. Reeves, D. Kontopantelis, E. Hiroeh, U., and M. Roland 2006. “Pay-for-Performance Programs in Family Practices in the United Kingdom.” *New England Journal of Medicine* 355, 375-384.

Dranove, D., Kessler, D., McClellan, M., and M. Satterthwaite. 2003. “Is More Information Better? The Effects of “Report Cards” on Health Care Providers.” *Journal of Political Economy* 111(3), 555-588.

Dranove, D. and G. Zhe Jin. 2010. “Quality Disclosure and Certification: Theory and Practice.” *Journal of Economic Literature* 48(4), 935–63.

Duckett, S., Daniels, S., Kamp, M., Stockwell, A., and G. Walker. 2008. “Pay for Performance in Australia: Queensland’s new Clinical Practice Improvement Payment.” *Journal of Health Services Research and Policy* 13, 174-177.

Dumont, E., Fortin, B., Jacquement, N., and B. Shearer. 2008. “Physicians' Multitasking and Incentives: Empirical Evidence from a Natural Experiment.” *Journal of Health Economics* 27, 1436-1450.

Ellis, R. and T.G. McGuire. 1986. “Provider Behavior under Prospective Reimbursement: Cost Sharing and Supply.” *Journal of Health Economics* 5, 129-151.

Engelmann, D. and M. Strobel. 2004. „Inequality aversion, efficiency, and maximin preferences in simple distribution experiments.” *American Economic Review* 94(4), 857–869.

Falk, A. and J.J. Heckman. 2009. “Lab experiments are a major source of knowledge in the social sciences.” CESifo working paper, No. 2894.

Fehr, E. and J.A. List. 2004. “The Hidden Costs and Returns of Incentives-Trust and Trustworthiness among CEOs.” *Journal of the European Economic Association* 2(5), 743-771.

Fischbacher, U., Kairies, N., and U. Stefani. 2013. “Equal Sharing, Accountability, and Productivity: An Experiment on Distributive Fairness in Complex Team Production Processes.” mimeo.

Flood, M.M. 1952. “Some Experimental Games.” Research Memorandum RM-789, RAND Corporation, June.

Frey, B. and F. Oberholzer-Gee. 1997. “The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out”, *American Economic Review* 87 (4), 746-755.

Gächter, S. and A. Riedl. 2005. “Moral property rights in bargaining with infeasible claims.” *Management Science* 51(2), 249–263.

Gantner, A., Güth, W., and M. Königstein. 2001. “Equitable choices in bargaining games with joint production.” *Journal of Economic Behavior and Organization* 46(2), 209–225.

Gené-Badia, J., Escaramis-Babiano, G., Sans-Corrales, M., Sampietro-Colom, L., Aguado-Menguy, F., Cabezas-Pena, C., and P. Gallo-Puelles. 2007. “Impact of economic incentives on quality of professional life and on end-user satisfaction in primary care.” *Health Policy* 80, 2–10.

Gneezy, U., and A. Rustichini. 2000a. "A Fine Is a Price." *Journal of Legal Studies* 29(1), 1-17.

Grabowski, D.C. and R.J. Town. 2011. “Does information matter? Competition, quality, and the impact of nursing home report cards.” *Health Services Research* 46, 1698–1719.

Gravelle, H., Sutton, M., and A. Ma. 2010. “Doctor behaviour under a pay for performance contract: treating, cheating and case finding?” *Economic Journal* 120, 129–156.

Grimm, V, Riedel, F., and E. Wolfstetter. 2003. “Low price equilibrium in multi-unit auctions: the GSM spectrum auction in Germany.” *International Journal of Industrial Organization* 21(10), 1557-1569.

Güth, W., Schmittberger, R., and B. Schwarze. 1982. “An experimental analysis of ultimatum bargaining.” *Journal of Economic Behavior and Organization* 3(4), 367–388.

Hall, Brian J., and Kevin J. Murphy. 2003. “The Trouble with Stock Options.” *Journal of Economic Perspectives* 17, 49–70.

Hennig-Schmidt, H., Selten, R., and D. Wiesen. 2011. “How Payment Systems Affect Physicians' Provision Behavior? An Experimental Investigation.” *Journal of Health Economics* 30(4), 637-646.

Hall, B.J. and K.J. Murphy. 2003. “The Trouble with Stock Options.” *Journal of Economic Perspectives* 17, 49–70.

Hannan, E.L., Kilburn, H., Racz, M., Shields, E., and M.R. Chassin. 1994. “Improving the Outcomes of Coronary Artery Bypass Surgery in New York State.” *Journal of the American Medical Association* 271(10), 761–66.

Holmstrom, B. 1982. “Moral Hazard in Teams.” *Bell Journal of Economics* 13, 324-40.

Holmstrom, B. and P. Milgrom. 1991. “Multi-Task Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job-Design.” *Journal of Law, Economics and Organization* 7, 24-52.

Iversen, T. and H. Lurås. 2000. “The effect of capitation on GPs' referral decisions.” *Health Economics* 9(3), 199-210

Kahn, L.M. 2000. “The Sports Business as a Labor Market Laboratory.” *Journal of Economic Perspectives* 14(3), 75-94.

Kairies, N. 2012. “P4P, Reputation and the Reduction of Costly Overprovision.” *Ruhr Economic Papers* No. 331.

Kairies, N. and M. Krieger 2013. “How do Non-Monetary Performance Incentives for Physicians Affect the Quality of Medical Care? A Laboratory Experiment.” *Ruhr Economic Papers* No. 414.

Karagözoğlu, E. and A. Riedl. 2011. “Information, uncertainty, and subjective entitlements in bargaining.” Working paper, revised version of IZA working paper #5079.

Kim, B.Y. 2010. “Quality Assessment and Pay for Performance in Korea.” HIRA, *Korean Health Insurance Review Agency*.

Konow, J. 1996. “A Positive Theory of Economic Fairness.” *Journal of Economic Behavior and Organization* 31(1), 13–35.

Konow, J. 2000. “Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions.” *American Economic Review* 90(4), 1072-1091.

Konow, J., Saijo, T., and K. Akai. 2009. “Morals and mores: Experimental evidence on equity and equality.” Mimeo, Loyola Marymount University.

Krasnik, A., Groenewegen, P.P. Pedersen, P.A. Scholten, P. Mooney, G. Gottschau, A. Flierman, H.A., and M.T. Damsgaard. 1990. “Changing Remuneration Systems: Effects on Activity in General Practice.” *British Medical Journal* 300, 1698-1701.

Levitt, S.D. and J.A. List. 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?.” *Journal of Economic Perspectives* 21(2), 153-174.

Levitt, S.D. and J.A. List, 2009. "Field experiments in economics: The past, the present, and the future." *European Economic Review* 53(1), 1-18.

Levitt, S. D. and S. J. Dubner. 2009. "Superfreakonomics." William Morrow, New York.

Marshall, M.N., Shekelle, P.G., Leatherman, S., and R.H. Brook. 2000. "The Public Release of Performance Data: What Do We Expect to Gain? A Review of the Evidence." *Journal of the American Medical Association*, 1866-1874.

Maynard, A. 2012. "The powers and pitfalls of payment for performance." *Health Economics* 21(1), 3-12.

McClellan, M. 2011. "Reforming payments to healthcare providers: The key to slowing healthcare cost growth while improving quality?" *Journal of Economic Perspectives* 25, 69-92.

Milgrom, P. and J. Roberts. 1988. "An Economic Approach to Influence in Organizations." *American Journal of Sociology* 94, 154-79.

Mosteller, F. and P. Nogee. 1951. "An Experimental Measurement of Utility." *Journal of Political Economy* 59, 371-404.

Mullen, K., Frank, R., and M. Rosenthal 2010. "Can You Get What You Pay For? Pay-For-Performance and the Quality of Healthcare Providers." *RAND Journal of Economics* 41(1), 64-91.

Muralidharan, K. and V. Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1), 39-77.

Nielsen, K. 1985. "Equality and Liberty." Totowa, NJ: Rowman and Allanheld.

Nozick, R. 1974. "Anarchy, State, and Utopia." New York: Basic Books.

Paarsch, H. and B. Shearer. 1999. "The Response of Worker Effort to Piece Rates: Evidence from the British Columbia Tree-Planting Industry." *Journal of Human Resources* 34 (4), 643-667.

Paarsch, H. and B. Shearer. 2000. "Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records." *International Economic Review* XLI, 59 –92.

Popper, K.R. 1934. "Logik der Forschung." J.C.B. Mohr J. C. B. Mohr (Paul Siebeck), Tübingen.

Popper, K.R. 1963. "Conjectures and Refutations: The Growth of Scientific Knowledge." Routledge Classics, London.

Prendergast, C. J. 1999. “The provision of incentives in firms.” *Journal of Economic Literature* 37, 7-63.

Perkins, R. and M. Seddon. 2006. “Quality improvement in New Zealand healthcare. Part 5: measurement for monitoring and controlling performance–the quest for external accountability.” *New Zealand Medical Journal* 119.

Rege, M. and K. Telle. 2004. “The Impact of Social Approval and Framing on Cooperation in Public Good Situations.” *Journal of Public Economics* 88, 1625-1644.

Rosenthal, G. E., Quinn, L., and D. L. Harper. 1997. “Declines in hospital mortality associated with a regional initiative to measure hospital performance” *American Journal of Medical Quality* 12(2), 103-112.

Rosenthal, M.B., Fernandopulle, R., Song, H.R., and B. Landon. 2004. “Paying for Quality: Providers’ Incentives for Quality Improvement, *Health Affairs* 23(2), 127-41.

Rosenthal, M.B., 2008. “Beyond Pay for performance – Emerging Models of Provider Payment Reform.” *New England Journal of Medicine* 359, 1197-1200.

Roth, A.E. (1995). “Introduction to experimental economics”, in The Handbook of Experimental Economics, ed. by J.H. Kagel and A.E. Roth, Princeton University Press, 3-110.

Roth, A.E., Sonmez, T., and M.U. Unver. 2004. “Kidney Exchange.” *Quarterly Journal of Economics* 119,457-488.

Roth, A.E., Sonmez, T. and M.U. Unver. 2005. “Pairwise Kidney Exchange.” *Journal of Economic Theory* 125, 151-188.

Scott, I. 2008. “Pay for Performance Programs in Australia: A Need for Guiding Principles.” *Australian Health Review* 32, 740–9.

Scott, A., Schurer, S., Jensen, P.H., and P. Sivey. 2009. “The Effects of an Incentive Program of Quality of Care in Diabetes Management.” *Health Economics* 9, 1091–108.

Siciliani, L. 2009. “Paying for Performance and Motivation Crowding Out.” *Economics Letters* 103, 68-71.

Sauermann, H. and R. Selten. 1960. “An Experiment in Oligopoly.” General Systems, Yearbook of the Society for General Research 5, 85-114.

Siegel, S. and L.E. Foraker. 1960.”Bargaining and Group Decision Making: Experiments in Bilateral Monopoly.” *American Economic Review* 64(5), 915-30.

von Neumann, J. and O. Morgenstern. 1944. “The Theory of Games and Economic Behavior.” Princeton: Princeton University Press.

Werner, R.M. and D.A. Asch. 2005. “The Unintended Consequences of Publicly Reporting Quality Information.” *Journal of the American Medical Association* 293(10), 1239-1244.

Werner, R.M., Konetzka, R.T., Stuart, E.A., Norton, E.C., Polsky, D., and J. Park. 2009. “The impact of public reporting on quality of postacute care.” *Health Services Research* 44(4), 1169–87.

Chapter 1

Equal Sharing, Accountability, and Productivity: An Experiment on Distributive Fairness in Complex Team Production Processes

Reference

Fischbacher, U., Kairies, N., and U. Stefani (2013): "Equal Sharing, Accountability, and Productivity: An Experiment on Distributive Fairness in Complex Team Production Processes."

Equal Sharing, Accountability, and Productivity: An Experiment on Distributive Fairness in Complex Team Production Processes

Abstract

We investigate the importance of different norms of distributive fairness in a team production process in which individual contributions to the joint output are not directly apparent. In particular, we investigate whether people prefer to share a joint output equally (egalitarianism), hold others responsible for their inputs (accountability principle), or reward others for their additional contributions to the joint output (libertarianism). The results of our experimental study indicate that people hold others responsible only for controllable factors, i.e., they reward others for their inputs. The predominance of the accountability principle is robust to the role of the person distributing the joint production payoff (neutral third party vs. proposers and responders in an Ultimatum Game) and to the framing of the production process (simultaneous vs. sequential). Marginal productivities are taken into account only when they are made readily apparent. Thus, our findings offer an explanation of why subjective performance measures are frequently regarded as unfair.

I. Introduction

In practice, situations in which members of a team or group must individually exert some effort or make relationship-specific investments in order to realize a joint output or profit are abundant. Real-world examples include work teams within a firm, business partnerships, corporate networks, joint ventures, strategic alliances of companies, research collaborations, and co-authored publications. In these situations, mutually agreed-upon rules determining the *ex post* allocation of the joint output to team members are an important mechanism to provide team members with sufficient (monetary) incentives *ex ante*. However, motivation and thus (non-monetary) incentives might be negatively affected if team members regard the resulting income distribution as unfair. In this context, the degree to which individual performance is observable not only determines the scope of performance measures available to create sufficient pecuniary incentives, but also influences the norms of distributive fairness prevalent within the team.

In situations in which *individual performance* is observable, one efficient compensation scheme is the implementation of an allocation rule that takes into account team members' contributions, effort levels, or inputs (*Alchian and Demsetz* (1972)). A typical example is a piece-rate compensation scheme, e.g., payment per unit of installed car windows (*Lazear* (2000)), payment per planted tree (*Paarsch and Shearer* (1999)), or payment per unit of harvested fruit (*Bandiera et al.* (2007)).

Usually, however, team production processes are more complex. For example, for the majority of workers, individual performance within the work team might be hard to observe (*Prendergast* (1999)), and marginal contributions to the joint output might be even more difficult to measure. In cases in which individual performance and marginal contributions are not verifiable, team members are frequently remunerated by means of a pre-defined fixed share of the *joint output* (i.e., by implementing some kind of "profit sharing"). Thus, the performance measure does not directly depend on individual contributions. An example is the joint venture (named *Caradigm*) formed by *Microsoft* and *General Electric* as a collaboration in the health care sector.[1] Combining *Microsoft's* IT expertise with the experience of *General Electric* in health care services allows both companies to generate new products, ultimately yielding additional profits for both. *Microsoft* contributes various patented software technologies (*Amalga*, *Vergence*) to the joint venture, while *General Electric* provides a huge clinical data-

[1] See the press releases that can be found at http://www.microsoft.com/en-us/news/press/2012/feb12/02-13HIMSPR.aspx (retrieved: 14.01.2013) and http://www.microsoft.com/en-us/news/Press/2012/Jun12/06-06CaradigmPR.aspx (retrieved: 14.01.2013).

base (*Qualibria*). However, the additional contributions of the two companies (e.g., human capital, existing customer relationships, or effort exerted) are hard to measure. As the joint venture is 50-50 by design, the two companies agree to equally share all future profits, independent of their individual contributions. However, one drawback of compensation schemes that do not account for individual performance is that they can lead to shirking and to inefficient effort levels in the sense of free-riding (*Alchian and Demsetz* (1972), *Baker et al.* (1988)).

One possible way to mitigate this incentive problem is to reward team members by means of *subjective performance measures* (*Rajan and Reichelstein* (2006)) that make use of partially non-verifiable information, e.g., the "value of the team member for the organization" or other proxies for non-observable individual inputs or marginal productivities. However, if team members have non-standard preferences, subjective performance evaluations can create new incentive problems (*Baker et al.* (1988)). One example is the evaluation process for junior faculty at research universities: Frequently, the evaluation scheme is not explicitly specified *ex ante*, and it probably does not entail a complete measurement of all the components that characterize a good scholar. Another recently discussed example that has attracted media attention is the compensation of top executives. In particular, empirical evidence shows that CEO pay and the company's contemporaneous performance are not closely linked (*Izan et al.* (1998), *Shiwakoti et al.* (2004), *Kubo* (2005), and *Doucouliagos et al.* (2007)). Case studies of *Bear Sterns* and *Lehman Brothers*, for instance, reveal that between 2000 and 2008, their top executives cashed out performance-based compensation of $1.4 billion and $1 billion, respectively. These amounts could not be clawed back when the companies collapsed. The stakeholders and owners of companies thus frequently question appropriateness – and fairness – of top management's pay. In April 2012, for example, stockholders accounting for 55% of *Citigroup's* shares did not approve the pay of the company's chief executive, Vikram Pendit. In the previous year, Pendit had earned a total of $15 million, even though the company's stock had halved in value and paid a dividend of only 1 cent per share.[2]

These examples vividly illustrate that understanding the norms of distributive fairness in complex team production processes is crucial for the design of suitable performance measures. Investigation of this problem raises the following question: In cases in which individual marginal contributions to the joint output are not directly obvious (e.g., in the presence of synergies or negative externalities), what allocation do team members regard as fair?

2 See http://www.ft.com/cms/s/0/6eddcc0e-8d63-11e1-b8b2-00144feab49a.html (retrieved: 13.01.2013).

The literature discusses a variety of fairness norms, which vary in the extent to which different factors constituting the production process are considered in allocating the joint output. More precisely, the elements of the production process might not be controllable by an individual to the same extent. Under the fairness ideal of *strict egalitarianism*, for example, team members should receive equal shares of the joint output (e.g., in cases in which a real-effort task is performed and there are no monetary effort costs) or equal incomes (in cases with monetary effort or investment costs). The *accountability principle* recommends that allocations only depend on (endogenous) factors that are completely under an individual's control (e.g., the effort level, the effort intensity, or the investment level chosen), whereas individual characteristics beyond an individual's control (e.g., talent, ability, or traits) are neglected. Under the fairness ideal of *liberal egalitarianism*, in contrast, both effort and individual characteristics affect distributional choices. Finally, the fairness ideal of *libertarianism* proposes that individuals' contributions to the joint output should determine their shares – i.e., in addition to effort levels and individual characteristics, exogenous factors such as marginal productivities should determine the allocation of the joint output.

In the present study, we focus on the importance of exogenous factors such as individual marginal productivities for the allocation of an output generated within a team. In our experiment, we investigate a situation in which two group members must answer a knowledge question to increase the joint output available for distribution. Varying the increase in the joint output resulting from each team member giving the correct answer allows us to investigate whether individual inputs (i.e., giving the right answer *per se*) or marginal productivities (i.e., the increase in Points realized by giving the correct answer) affect distributional choices. For each treatment, we observe distributional choices for nine different situations.

In one series of experiments, a third party outside the team allocated the output to the two subjects who jointly generated the output within their team (*Third Party Treatments*). Since the third parties' payoffs were not affected by their distributional choices, we should be able to observe the relative importance of fairness norms in an unbiased fashion.

In the *simultaneous framing* condition of the Third Party Treatment (TPT), we showed our participants the payoff matrices resulting from the nine different situations. Thus, the individual marginal productivities realizable within the group were observable. However, since our design included cases with synergies and negative externalities, the reference point for the distribution of the joint output was not clear. Our results indicate that libertarianism does not explain the data, i.e., that the allocation of the joint output did not depend on exogenous marginal productivities. Instead, group members rewarded endogenous inputs and allocated more Points to the subject who knew the correct answer to the knowledge question. Consequently, both team members received an equal share if neither or both subjects gave the correct answer. If only one subject gave the right answer, this participant received a share of approximately two-thirds of the joint output.

In the *sequential framing* condition of the Third Party Treatment (TPTSQM), we explicitly informed our participants about the *additional increase* in the joint output they could realize by giving the correct answer. Although the TPT and the TPTSQM conditions are economically completely equivalent, we found that subjects considered marginal productivities only in the TPTSQM condition, in which output increases were made overtly apparent. One reason for this could be that the information about marginal productivities created personal entitlements to claim a share based on the increase in points a subject could realize.

However, the difference between the TPT and the TPTSQM lies not only in the explicitness of the marginal productivities, but also in the sequential framing necessary to inform subjects about the output increase that would result from a correct answer to our knowledge question. In order to rule out the possibility that it was the sequential framing rather than the obviousness of the incremental increase in the joint output that caused the treatment effect we observed, we implemented a second sequential framing condition of the Third Party Treatment, in which we informed our subjects about the possible *joint output payoffs* (instead of the resulting payoff increases) that team members could realize by giving the correct answer (TPTSQNM). Since our results from the TPTSQNM treatment are qualitatively similar to those from the TPT, sequential framing does not seem to be the reason for the shift from the accountability principle to libertarianism.

In reality, situations in which a non-involved person outside the team distributes the joint payoff occur rather infrequently. Thus, we also investigated the more realistic case in which group members allocated an output they had previously generated within their groups, testing whether their distributive choices would be different from those of neutral third parties. Using the rules of the *Ultimatum Game* to allocate the joint output (UGT), we found that subjects in the UGT, the TPT, and the TPTSQNM in principle operated on the basis of the same fairness norms. Subjects in the UGT, however, were somehow “less extreme”, in the sense that they did not reward the participant who knew the correct answer to the same extent that subjects in the TPT did. Moreover, proposers and responders within the UGT shared the same fairness norm. Thus, the prevalence of the accountability principle seems not to be affected by the role of the person allocating the joint output (third party vs. implicated parties), by the connection between subjects' distributive choices and their own payoffs, or by aspects of strategic fairness.

This paper proceeds as follows: Section II reviews the related literature. Section III presents the design of our experiment, and Section IV derives the predictions for the fairness norms we consider. Section V presents the results of our experimental study. Section VI summarizes our key findings and offers concluding remarks.

II. Related Literature

Early experiments investigating norms of distributive fairness assumed the distributable amount to be "manna from heaven" (*Güth et al.* (1982)). In real-life situations, however, frequently the output available for allocation must first be produced. Experimental evidence shows that in such cases, the prevalence of norms of distributive fairness is highly context-dependent. In the following section, we discuss some of the major issues.

II.1 Type of Input Used to Generate the Joint Output

First, the relative importance of norms of distributive fairness depends on the type of input necessary to produce the output. In particular, some experiments implement a production function with monetary investments, while others use a real-effort task.

Findings from experiments using *monetary investments* indicate that the majority of subjects prefer equal shares. However, a considerable percentage of subjects also acknowledge individual contributions, which can be measured in terms of the input contributed or in terms of the marginal productivity.

Gantner et al. (2001) suggest that the joint output itself or the surplus (joint output less investment costs) could be considered as output measures, whereas the investment costs or the individual contribution to the joint output (investment costs times individual marginal productivities) could be regarded as input measures. *Gantner et al.* (2001) find that, within an Ultimatum Game, 32% of the demand choices were explainable with equal sharing, 12% coincided with liberal egalitarianism, and 17% corresponded to libertarianism.[3]

Cappelen et al. (2007) explicitly differentiate between inputs that a subject can control (e.g., the ratio between individual investments and total investments) and factors beyond the participant's control (e.g., individual investments multiplied by the exogenously given productivity

[3] Similar experiments using an Ultimatum Game with additive and asymmetric advance production have been conducted by *Königstein* (2000) and *Brandstätter and Königstein* (2001).

parameter). In their Dictator Game with a preceding phase of joint production, *Cappelen et al.* (2007) find that 43% of the subjects' allocation choices were in line with strict egalitarianism and 40% with liberal egalitarianism, with only 17% following libertarianism.

Implementing monetary efforts such as those in *Gantner et al.* (2001) and *Cappelen et al.* (2007), however, complicates the disentanglement of preferences for equal payoffs from preferences for the acknowledgment of individual contributions or performance in a general sense. In this regard, *Gächter and Riedl* (2005) have shown that "moral property rights" can significantly shape bargaining behavior. Along similar lines, evidence from experiments with joint production in the form of *real-effort tasks* is less supportive of equal sharing and indicates increased acknowledgement of individual inputs.

In this context, *Konow* (2000) describes an experiment in which subjects performed a real-effort task (folding letters) and distributed the joint output according to the rules of the Dictator Game. In accordance with the accountability principle, the dictators' shares varied in direct proportion to their entitlements, which in turn depended on the number of letters folded. However, *Konow* (2000) focuses on a single fairness norm.

Cappelen et al. (2011) study the relative importance of needs, entitlements, and nationality in people's fairness norms. Joint production was implemented with a real-effort task (typing words), and distribution was accomplished using a Dictator Game. The experiment was conducted with participants in two poor (Uganda and Tanzania) and two rich countries (Germany and Norway). *Cappelen et al.* (2011) find that liberal egalitarianism was the most prevalent distributive norm among rich participants, at 42%. Libertarians were estimated at 34%, while the share of egalitarians was only 23.8%. The focus of *Cappelen et al.* (2011), however, is on explaining how people distribute when different moral motives are at play.

Cappelen et al. (2010) disentangle liberal egalitarianism by further analyzing the factors for which people are held responsible. In their experiment, individual contributions resulted from subjects' inputs (i.e., the performance of a real-effort task) or choices (i.e., time spent working) and a randomly assigned value-per-input unit. In line with the previous results, *Cappelen et al.* (2010) find that liberal egalitarianism is the most prevalent fairness norm, in particular in the form of holding people responsible for their inputs. The percentages of subjects who allocated according to the accountability principle, libertarians, strict egalitarians, and those holding others responsible for their choices were 42%, 28%, 24%, and 7%, respectively.

In summary, equal sharing tends to be the most prevalent norm when monetary investments are necessary to generate an output, whereas fairness based on accountability gains in importance when real-effort tasks are performed. In our experiment, subjects had to perform a real-effort task. Thus, we can differentiate between the preference for taking into account inputs *per se* and the preference for equal income distributions.

II.2 Role of the Person Allocating the Joint Output

The prevalence of fairness norms can also be driven by the role of the person allocating the joint output. In this context, *Konow et al.* (2009) compare the distributive choices of stakeholders allocating a previously produced joint output according to the rules of the Dictator Game with those of neutral third parties. *Konow et al.* (2009) find that third parties use fairness norms that account for individual contributions more often than stakeholders do, and that only lifting anonymity shifts the balance towards equality.

In our experiment, we ran sessions in which subjects distributed an output that they had previously generated within their group (Ultimatum Game Treatments) as well as sessions in which neutral third parties allocated the output produced by a group they did not belong to (Third Party Treatments). Thus, we can address the question of whether the prevalence of fairness norms depends on the role of the person allocating the output, including issues such as potential self-serving biases.

II.3 Complexity of the Production Process – Additivity and Observability

Distributive fairness can also be influenced by increased complexity in the production process. First, joint production must not necessarily be additive; it can also lead to negative externalities or to synergies. *Croson et al.* (2004), for example, describe a number of mergers and acquisitions that resulted in synergies and in negative externalities.

Gächter and Riedl (2005) experimentally analyze the distribution of a joint output in the presence of negative externalities. Subjects completed a real-effort task and then received a claim that was dependent on their relative performance in a group of two. In some of the cases, however, the sum of the proposed claims exceeded the distributable amount. *Gächter and Riedl* (2005) find that also in these cases, claims resulting from relative performance served as focal points.

Second, individual performance, especially in the form of marginal productivities, is often difficult to observe (*Prendergast* (1999)). *Karagözoğlu and Riedl* (2011) address this issue by extending the framework presented by *Gächter and Riedl* (2005) to uncertainty about relative performance, demonstrating that receiving information about individual performance influences entitlements and the bargaining process.

Although *Gächter and Riedl* (2005) and *Karagözoğlu and Riedl* (2011) show that entitlements and uncertainty about relative performance influence the allocation chosen, they do not investigate the prevalence of distributive fairness norms. Since the subjects in their experiments received information only about their relative performance, performance in the form of inputs or marginal productivities was completely unobservable. However, this information can significantly shape entitlements, distributive fairness, and allocations.

In our experiment, we take *additive marginal payoffs*, *negative externalities*, and *synergies* into account. This enables us to investigate the importance of fairness norms across a wide range of joint production outcomes. Moreover, we implemented a *simultaneous framing* condition by showing participants the resulting payoff matrices (TPT), as well as *sequential framing* conditions, in which we explicitly informed participants about the additional output increase (TPTSQM) and the resulting joint output (TPTSQNM), respectively, that team members could realize. We can therefore address the question of whether the prevalence of fairness norms depends on the noticeability of individual marginal productivities.

III. Experimental Design

III.1 General Description of Our Experiment

The focus of our experiment is on the distribution of a joint output that has been produced within a group of two subjects. We randomly formed groups of two participants, one in the role of subject A and one in the role of subject B, and asked both group members to answer the same multiple-choice knowledge question. The output actually realized was determined by whether neither, one, or both subjects within a group knew the correct answer to this question; the joint production payoff increased when one of the group members answered the question correctly, and it was highest when both subjects knew the correct answer. We kept the joint output in the symmetric cases (in which neither or both subjects within the group knew the correct answer) constant at 20 Points and at 100 Points, respectively, and varied the output for the asymmetric cases (in which only one of the participants gave the right answer) to 30, 60, and 90 Points. As illustrated in Table 1, there are thus nine different joint production payoff matrices, which we denote by $JP_m = \begin{bmatrix} 20 & A_w B_r \\ A_r B_w & 100 \end{bmatrix}$ (with $m = 1, ..., 9$). $A_w B_r$ $(A_r B_w)$ refers to the asymmetric case in which subject A (subject B) gives the wrong answer to the question, while subject B (subject A) gives the right answer.

Insert Table 1 about here

Since our variation of the joint production payoffs allows for a clear differentiation between varieties of marginal productivities, we can test whether the (exogenously given) increase in the joint output resulting from one of the group members answering the question correctly affects distributional choices. In our experiment, subjects made decisions on all nine payoff matrices, i.e., not only could they clearly observe the respective individual marginal productivities of the group members, but they could also recognize the differences regarding marginal productivities between matrices.

Note that the marginal production payoffs are additive only in matrices $JP_3 = \begin{bmatrix} 20 & 90 \\ 30 & 100 \end{bmatrix}$, $JP_5 = \begin{bmatrix} 20 & 60 \\ 60 & 100 \end{bmatrix}$, and $JP_7 = \begin{bmatrix} 20 & 30 \\ 90 & 100 \end{bmatrix}$. In these matrices, the marginal productivities in all three cases $A_r B_w$, $A_w B_r$, and $A_r B_r$ are evident: When participants give the correct answer, the effects on the size of the distributable output are precisely determinable. Matrices $JP_1 = \begin{bmatrix} 20 & 30 \\ 30 & 100 \end{bmatrix}$, $JP_2 = \begin{bmatrix} 20 & 60 \\ 30 & 100 \end{bmatrix}$, and $JP_4 = \begin{bmatrix} 20 & 30 \\ 60 & 100 \end{bmatrix}$, in contrast, are super-additive, i.e., $100 > 20 + (A_w B_r - 20) + (A_r B_w - 20)$. Since the group members' marginal productivities add up to less than 80 Points, subjects can realize synergy effects if they both know the right answer. It is thus not apparent which of the group members played the greater role in earning the joint output of 100 Points. However, when only one group member answers correctly, the resulting effects on the joint output are again precisely attributable. Matrices $JP_6 = \begin{bmatrix} 20 & 90 \\ 60 & 100 \end{bmatrix}$, $JP_8 = \begin{bmatrix} 20 & 60 \\ 90 & 100 \end{bmatrix}$, and $JP_9 = \begin{bmatrix} 20 & 90 \\ 90 & 100 \end{bmatrix}$ are sub-additive, since the sum of the individual marginal productivities exceeds 80 Points. Given that one of the subjects knows the right answer, the other group member thus creates negative externalities on the joint production payoff by also giving the correct answer. We included the non-additive matrices in our design to investigate whether (and how) synergy effects and negative externalities in the case in which both subjects know the right answer affect distributional choices. In particular, we can test whether marginal productivities are more relevant for the additive matrices (in which marginal productivities are more apparent) than for the non-additive matrices.

For the allocation stage of our experiment, we implemented variants of two basic treatment conditions (Third Party Treatment and Ultimatum Game Treatment). We conducted only one treatment per session.

In the *Third Party Treatment* (*TPT*), each subject assumed the role of a third party in deciding on the allocation of a joint production payoff that *two other subjects* had created within their group by answering their respective knowledge questions. Since subjects in the role of the third party made decisions regarding the allocation of an output that an anonymous group had realized, and because their distributional choices did not affect their own payoffs, self-interest should not have impacted their judgments. Thus, the TPT enables us to observe norms of distributive fairness that are not distorted by self-serving behavior or self-centered fairness perceptions.

In addition, we ran a Third Party Treatment framed in a way that made marginal productivities even more apparent than showing subjects the matrices displayed in Table 1. Specifically, we chose a sequential presentation format in which we informed participants about the additional output increase they could realize by giving the correct answer (TPTSQM). Comparing the TPT with the TPTSQM allows us to investigate whether an explicit framing of marginal productivities affects the relative importance of the libertarian fairness ideal. We also conducted another Third Party Treatment with sequential framing, but without explicitly informing participants about the marginal productivities of the team members (TPTSQNM). In the TPTSQNM, we showed participants the joint output resulting from neither / one / both subjects knowing the correct answer. Consequently, we can isolate the effects of the sequential framing from those resulting from the apparentness of marginal productivities.

Situations in which an "impartial arbitrator" allocates a distributable amount to the individuals involved in the production of the joint output (as in our TPT) are, however, uncommon in actual practice. Clearly more realistic are situations in which group members must come to an agreement on how to apportion a collaboratively generated output. In our *Ultimatum Game Treatment* (UGT), subjects thus had to decide on the allocation of a joint production payoff that they had created *within their own group* according to the rules of the Ultimatum Game (*Güth et al.* (1982)). By comparing data from the TPT with data from the UGT, we can investigate whether the fairness norms we observe for the TPT also apply in a more realistic situation in which implicated parties with asymmetric bargaining power distribute a jointly generated output. Subjects in the UGT might feel entitled to claim a certain share of the output, in particular in a situation in which only one participant has answered the knowledge question correctly (or when the marginal productivity of one of the participants is comparatively large). Moreover, subjects, as implicated parties, might fall prey to a potential self-serving bias in making their fairness judgments (*Babcock et al.* (1995), *Babcock and Loewenstein* (1997), *Charness and Haruvy* (2000), *Konow* (2000), *Gächter and Riedl* (2005), and *Konow* (2009);

for an overview, see *Konow* (2003)). This suggests that "individuals (might) subconsciously alter their fundamental views about what is fair or right in a way that benefits their interests" (*Dahl and Ransom* (1999), p. 703). Comparing third parties' decisions with those of subjects that have a personal stake in the outcome of both the production stage and the allocation stage thus enables us to test whether there are differences in fairness perceptions between the two methods of allocation.

In the UGT, we implemented a role reversal (i.e., half of the subjects first made their decisions in the role of the proposer (responder) and then in the role of the responder (proposer)). By contrasting proposers' offers with responders' acceptance thresholds (within-subject comparison), we can test whether individual perceptions of distributive fairness are sensitive to the subject's role. To rule out the possibility that the role reversal could affect distributional choices (e.g., that subjects' offers and minimum claims would convert to 50% due to the role reversal), we conducted an additional session in which subjects distributed a joint production payoff that they had generated within their own group according to the rules of the Ultimatum Game, but only in one role, either as the proposer or the responder (UGT1R).

Using ORSEE (*Greiner* (2004)), we recruited students from the University and the University of Applied Sciences in the same town for our experiment. We programmed and conducted the experiment using the z-Tree software (*Fischbacher* (2007)). Each subject participated in only one of our sessions.

After all of the subjects had entered the laboratory, they were seated at individual computer workstations. Participants received instructions for their respective roles, which we asked them to read carefully. Subjects then answered some control questions. The experiment began after we had orally summarized the instructions. We gave this detailed vocal summary to ensure that all participants understood the rules of the game and knew that in the Ultimatum Game Treatments (UGT and UGT1R) the instructions for proposers and responders followed the same set of rules.

At the end of the experiment, we converted the Points that subjects had earned into money (one Point was equivalent to 0.25 € ($ 0.40)) and paid out this amount anonymously, together with the 5 € ($7.90) participation fee. Table 2 shows the number of subjects per treatment, as well as the average payments that participants earned for a session of approximately 50 minutes. The compensation that participants received is well above the hourly rate for student jobs.

Insert Table 2 about here

III.2 Detailed Information on the Experimental Design

For all of our treatments, we implemented the strategy method (*Selten* (1967)) for the allocation stage. For each of the nine matrices specified in Table 1, subjects had to announce their preferred allocation of the joint production payoffs to group members *before* the knowledge question was asked.

In the TPT, TPTSQM, and TPTSQNM conditions, all subjects first decided as a third party on the distribution of joint production payoffs to an anonymous group of two members in the roles of A and B. Third parties had to make their distributional choices before they were randomly matched with the group whose payoff they decided. Note that third parties always allocated an output that they had not themselves created.

After subjects in the role of the third party had made their decisions, the experimental software randomly determined two of the nine matrices as payoff-relevant. For each of these matrices, we randomly built groups of two additional subjects from the subject pool we had recruited for the session, assigning them the roles of A and B, respectively. Thus, we implemented a role reversal, in the sense that every subject was now in the role of A or B, but always within a group that was different from the one the subject had decided for as a third party. Subjects A and B then realized joint production as follows.

All participants saw six multiple-choice questions from a knowledge quiz on their computer screens. The experimental software randomly chose one participant, who rolled a die to determine which one of these questions would be relevant for the first matrix. All participants had to answer the question that this subject had randomly chosen. Ensuring that questions would not be repeated if the same number came up twice the same procedure was carried out for the second matrix. The correctness of the answers of the two subjects A and B within a group determined the joint output realized (i.e., a specific cell of the randomly chosen matrix). When participants were answering the question, they knew which matrix applied, but we had not yet informed them with whom they were paired in a group and whether this subject had answered the question correctly. Subjects did not have an incentive to answer strategically, since their payoffs were always increased by answering correctly.

After all of the subjects had answered the knowledge question, each group was matched with a third party outside the group. One of the selected matrices determined a subject's payoff in role A, the other the payoff in role B. Thus, the distributional choices made by subjects in the role of the third party never affected their own compensation earned within the experiment. The selection of matrices, assignment of roles, building of groups, and matching were completely random. However, the allocation of the joint production payoff was carried out exactly as the third party the group was matched with had previously decided.

For the TPTSQM and the TPTSQNM, we implemented the same procedure as with the TPT, except for the framing. To make the increase in the joint production payoff that could be realized by giving the correct answer (i.e., the individual marginal productivities) more apparent in the TPTSQM, we presented the decisions in the TPTSQM sequentially. As an example, consider the joint production payoff matrix $JP_8 = \begin{bmatrix} 20 & 60 \\ 90 & 100 \end{bmatrix}$: We started with 20 Points and informed subjects on their computer screens that the joint production payoff would not increase further if A and B both gave the wrong answer. In the next line on the screen, we explicitly informed subjects that the output would increase by 40 Points if B (but not A) knew the correct answer. The following line stated that the joint output, starting from the baseline of 20 Points, would increase by 70 Points if A (but not B) answered the question correctly. The last line told subjects that the output would go up by 70 Points if A was right and by an additional 10 Points if B was also right. Thus, the framing in the TPTSQM implied that A was the first mover in creating the joint production payoff of 100 Points. In the TPTSQNM, we implemented the same sequential framing with A as the first mover. In contrast to the TPTSQM, however, we informed subjects in the TPTSQNM about the joint output that could be realized if neither A nor B / only B / only A / both A and B answered the question correctly.

In the UGT and the UGT1R, the two subjects within a group had to agree on how to distribute the joint production payoffs realized within their group according to the rules of the Ultimatum Game: Applying the strategy method, proposers (subjects A) announced the share of the joint payoff they wanted to offer to the responder within their group (subject B), and subjects B simultaneously announced the minimum thresholds they set for the acceptance of the offer from subject A. Should B's threshold be lower than A's offer, the subjects reached an agreement, and A's offer was carried out. If, however, B's threshold exceeded A's offer, neither subject received anything.

For each of the nine matrices specified in Table 1, subjects within a group in the UGT had to decide as a proposer as well as a responder. More precisely, in the UGT, we assigned half of the subjects to subgroup I and the other half to subgroup II. In subgroup I (II), subjects first decided as a proposer (responder) and then as a responder (proposer). Thus, for each of the nine matrices, we observed decisions as a proposer *and* as a responder from each subject. In the UGT1R, subjects did not change their roles, i.e., for each subject in the UGT1R, we only observed decisions as a proposer *or* as a responder.

After all of the participants in the UGT and the UGT1R had made their decisions, we again randomly selected two matrices to determine actual payoffs. Realization of the joint production payoff was carried out exactly as described for the Third Party Treatments. In the UGT (UGT1R), we randomly built groups of two participants from different subgroups (of subjects A and B) for each selected matrix, and the decisions made by the subjects within these groups were payoff-relevant. For each subject, one of the selected matrices was relevant for the decision as a proposer, the other matrix for the decision as a responder.

IV. Predictions

In the following section, we derive predictions with regard to the distribution of the joint production payoffs as defined in Table 1. Provided that subjects adhere to one of the fairness ideals we investigate, we should observe behavior in line with these predictions, in particular for subjects in the Third Party Treatments.

In the Ultimatum Game Treatments, in addition to a subject's individual fairness preference, other considerations could be relevant: First, proposers might care about fairness not only because they are interested in the responder's well-being ("intrinsic fairness"), but also because they fear the rejection of unfair offers ("strategic fairness"; *Mitzkewitz and Nagel* (1993), *Hoffman et al.* (1994), *Pillutla and Murnighan* (1995), *Straub and Murnighan* (1995), *Croson* (1996), *Kagel et al.* (1996), *Rapoport and Sundali* (1996), *Van Dijk et al.* (2004)). Strategic fairness implies that even if low offers were consistent with the proposer's "intrinsic fairness norm", we should not expect to observe comparatively low offers in the Ultimatum Game Treatments, since low offers are frequently rejected. Because proposers in the Ultimatum Game Treatments have incomplete information regarding the responders' fairness norms, proposers might regard an equal split as an offer that would minimize the risk of rejection. Second, since subjects' decisions in the Ultimatum Game directly affect their own payoffs, we should not expect to observe very high offers in the Ultimatum Game Treatments. As a consequence, offers in the UGT and in the UGT1R should be less widely scattered than allocations in the TPT, TPTSQM, and TPTSQNM. Moreover, minimum acceptance thresholds in the UGT and in the UGT1R are predicted to be lower than allocations in the TPT, TPTSQM, and TPTSQNM, because higher thresholds correspond with a higher incidence of bargaining failures (*Oosterbeek et al.* (2004)).

IV.1 Equal Sharing (Strict Egalitarianism)

If group members distribute joint payoffs according to the *strict egalitarian* fairness ideal, A and B should each receive a share of 50% of the joint output, irrespective of who has answered the question correctly and has thus increased the distributable endowment. Note that strict egalitarian fairness and equal sharing as modeled in *Fehr and Schmidt* (1999) and *Bolton and Ockenfels* (2000) lead to identical predictions within our experiment, because performing a real-effort task (i.e., giving an answer to our knowledge quiz) does not incur monetary effort costs.

IV.2 Fairness Based on Accountability

Fairness based on *individual accountability* presumes that knowing the answer to our knowledge question creates mutually accepted moral property rights to claim a certain share. Thus, participants should be rewarded simply for knowing the right answer, irrespective of the absolute amount of the output increase they were able to realize. For the symmetric cases in which both or neither of the group members knew the right answer, we again predict that subjects will distribute the joint production payoff equally, since both subjects within the group created the distributable endowment in equal measure. For the asymmetric cases, however, we predict that the subject who knew the correct answer will receive a larger share α of the joint production payoff than the subject who was wrong, i.e., $\alpha \geq 0.5$. If subjects somehow reward group members who give the wrong answer for participating in the experiment and for trying to answer the knowledge question, we should observe $\alpha < 1$. If subjects adhere to individual accountability, the share α that the subject who is responsible for increasing the joint payoff receives is predicted to be constant across payoff matrices, i.e., independent of the marginal productivities realized by knowing the right answer. Since we cannot distinguish talent or ability from effort in our design, we refrain from making predictions regarding the distributive choices of liberal egalitarians.

IV.3 Fairness Based on Individual Marginal Productivities (Libertarianism)

If subjects value the *libertarian fairness* ideal, their allocation should take into account incremental changes in the distributable output arising from the correctness of their answers. For all of our matrices, the incremental output effect resulting from only one group member answering the question correctly (i.e., for cells $A_r B_w$ and $A_w B_r$) is apparent. For our additive matrices (JP_3, JP_5, JP_7), marginal productivities are clearly attributable to subjects A and B, even for cell $A_r B_r$. For the non-additive matrices, in contrast, it is not clear which of the subjects had greater "responsibility" for creating the joint payoff in $A_r B_r$ (i.e., who the first mover was). Moreover, it is not obvious whether subjects are focused on gains or on losses, i.e., whether the joint production payoff of 20 Points or of 100 Points serves as the reference point. For both scenarios of reference points, we calculate Shapley Values (*Shapley* (1953)) to specify the average marginal contribution (marginal loss) each subject realizes by entering (leaving) the coalition. In our design, we assume a coalition to be the situation in which both subjects answer the question correctly. Since the Shapley Value takes into account the possibility that subjects A and B can each join the coalition as either a first player or a second player, it is a concept ideally suited for predicting shares when subjects adhere to the libertarian fairness ideal.

If the *focus is on gains* (i.e., the reference point is that neither of the subjects knows the right answer), subjects start from a joint production payoff of 20 Points. We assume that if both subjects give the wrong answer, 20 Points will be distributed equally. Consider the payoff matrix $JP_3 = \begin{bmatrix} 20 & 90 \\ 30 & 100 \end{bmatrix}$, for which marginal productivities are additive. In the asymmetric case $A_w B_r$ (90 Points), B's marginal productivity is 70 Points. Hence, B should receive 70 Points more than in the reference situation. Thus, we predict an allocation of 10 Points for A and 80 Points for B. If, in contrast, A answers the knowledge question correctly and B is wrong (30 Points), A's marginal productivity is 10 Points. Hence, we predict an allocation of 20 Points for A and 10 Points for B. For the symmetric case $A_r B_r$, B's Shapley Value, S_B^{Gain}, can be derived by dividing the sum of B's marginal contributions when B enters the coalition as a first player and as a second player by two. In our example, B entering the coalition as the first player would be a move from the symmetric case $A_w B_w$ (20 Points) to the asymmetric case $A_w B_r$ (90 Points). Thus, B's marginal contribution is 70 Points. Entering the coalition as the second player would represent a move from the asymmetric case $A_r B_w$ (30 Points) to the symmetric case $A_r B_r$ (100 Points), i.e., B's marginal contribution is again 70 Points. The Shapley Value for B is thus given by $S_B^{Gain} = \frac{70+70}{2} = 70$ Points; analogously, the Shapley Value for A is $S_A^{Gain} = \frac{10+10}{2} = 10$ Points. An example in which players' marginal contributions are not identical for the two types of move is provided by matrix $JP_2 = \begin{bmatrix} 20 & 60 \\ 30 & 100 \end{bmatrix}$.

Here, the Shapley Values are $S_A^{Gain} = \frac{10+40}{2} = 25$ Points and $S_B^{Gain} = \frac{40+70}{2} = 55$ Points. The individual Shapley Values (see Table 3) add up to 80 Points in each of the output matrices, which is exactly the total contribution realized when both group members answer the question correctly (100 Points) rather than incorrectly (20 Points).

Insert Table 3 about here

Table 4 shows the corresponding shares that B should receive (1) if the payoff $A_w B_w = 20$ is split equally, (2) if, in addition, individual marginal productivities in the asymmetric cases $A_w B_r$ and $A_r B_w$ are enforced, and (3) if shares of $\alpha_A = \frac{S_A^{Gain} + 10}{100}$ and $\alpha_B = \frac{S_B^{Gain} + 10}{100}$, respectively, are realized in the symmetric case $A_r B_r = 100$. Note that the share of the subject who answers correctly increases with the marginal productivity. We predict shares for B, α_B, to increase with the output resulting from situation $A_w B_r$ and to decrease with the output resulting from $A_r B_w$: The more A (B) contributes to the joint production payoff by knowing the right answer, the lower (higher) the share for B will be. In particular, shares should not depend on the fact that one of the subjects knew the right answer *per se*, but only on the increase in the distributable output realized by giving the right answer.

Insert Table 4 about here

Note that we have framed the TPTSQM and the TPTSQNM as situations in which subjects focus on gains and player A is the first mover. In analyzing these treatments, we have therefore calculated the variable *Shapley_Gain_SQ* to take into account the fact that A always enters the coalition as the first player.

If, however, subjects *focus on losses*, the reference point is the case in which both subjects know the right answer; again, we assume that subjects will share the resulting 100 Points equally in this case. For the payoff matrix $JP_3 = \begin{bmatrix} 20 & 90 \\ 30 & 100 \end{bmatrix}$, A (B) squanders 10 Points (70 Points) by not knowing the correct answer. Thus, these losses should be subtracted from the 50 Points that A and B would receive in the reference situation. Therefore, we predict $A_w B_r$

(90 Points) to be allocated at 40 Points for A and 50 Points for B. For the case A_rB_w (30 Points), A should receive 30 Points and B 0 Points, since the experimental design precludes negative shares. For the symmetric case A_wB_w (20 Points), we account for the sequential form of destroying Points by not knowing the right answer by again deriving Shapley Values. Since the Shapley Values listed in Table 3 do not depend on the reference point, A (B) should receive a share of $\alpha_A = min\left(\frac{50 - S_A^{Loss}}{20}; 1\right) = 1$ $\left(\alpha_B = max\left(0; \frac{50 - S_B^{Loss}}{20}\right) = 0\right)$ in the situation A_wB_w. Table 5 shows the predictions regarding shares for B when subjects adhere to libertarian fairness and focus on losses. In the table, we can see that shares for B should increase with the output resulting from A_rB_w and decrease with A_wB_r, i.e., the greater the amount of the joint production payoff that A (B) destroys by not knowing the right answer, the higher (lower) B's share will be. Again, shares should only depend on the marginal losses realized by not knowing the right answer, not on failing to answer the question correctly *per se*.

Insert Table 5 about here

Note that the TPTSQM and the TPTSQNM have been framed in terms of additional gains, so we do not expect a focus on losses to be behaviorally relevant in this treatment.

V. Results

V.1 Descriptive Statistics

V.1.a Result 1

In the symmetric case in which both subjects were wrong (A_wB_w), the joint production payoff of 20 Points was divided equally among group members.

Table 6 displays the shares that subjects in the Third Party Treatments allocated to subjects B, as well as the shares that proposers in the Ultimatum Game Treatments offered to the responders within their groups (averaged across all nine payoff matrices).

Insert Table 6 about here

Table 6 illustrates that in the symmetric case in which neither of the subjects within a group knew the correct answer, subjects, on average, shared the joint production payoff of 20 Points equally: When both subjects answered the question incorrectly, third parties in the TPT, TPTSQM, and TPTSQNM allocated approximately 50%, 50%, and 51% of the joint production payoff to subject B, and proposers in the UGT (UGT1R) offered responders 49% (48%) of the 20 Points.

Figure **1** shows the frequencies of shares allocated to subjects B in percent when both group members gave the wrong answer (averaged over all nine matrices). In line with *Result 1*, almost all third parties and the overwhelming majority of proposers distributed half of the joint production payoff to subject B.

Insert Figure 1 about here

Table 7 shows the allocations that third parties suggested for subject B, as well as proposers' offers and responders' minimum thresholds (averaged by payoff matrix). For the symmetric case A_wB_w, a two-sided t-test (a Wilcoxon signed-rank test) comparing allocations / offers to subjects B across treatment conditions reveals that only seven (eight) out of 90 situations were statistically different from the other situations at conventional levels of significance. Thus, the equal distribution of the 20 Points was unaffected by the treatment condition.

If we take data for the symmetric case A_wB_w and regress shares allocated to subjects B on dummy variables for the nine different payoff matrices, we do not find evidence of significant differences across payoff matrices within any of our five treatment conditions (only the allocation for A_wB_w in matrix JP_2 of the TPT was significantly different). Unsurprisingly, the payoff matrix did not affect the allocation of the 20 Points resulting from the fact that neither A nor B answered the question correctly.

Insert Table 7 about here

The equal allocation of 20 Points is in line not only with the predictions of fairness based on accountability (or liberal egalitarianism), but also with strict egalitarianism, since both group members created the distributable output in equal measure and did not have to incur monetary effort costs. If people focus on gains, the libertarian fairness ideal also predicts shares of 0.5 for all matrices. If people focus on losses, however, libertarian fairness predicts shares of 0.5 only for matrices JP_1, JP_5, and JP_9. Our data, in contrast, show that in all of our treatments, the shares allocated to B in the symmetric case A_wB_w were approximately 50%, even for the matrices in which libertarian fairness predicts shares other than 0.5 (indicated by the variables *Shapley_Loss* defined in Table 5). Thus, subjects did not seem to take marginal productivities into account.

V.1.b Result 2

In the asymmetric cases in which only one subject was right (A_rB_w and A_wB_r), the subject who answered the question correctly received more than half of the joint production payoff.

Table 6 also confirms that in the asymmetric cases (only one subject within a group knew the correct answer), group members, on average, distributed more than half of the joint production payoff to the subject who knew the correct answer to the question.

Third parties in the TPT, TPTSQM, and TPTSQNM, on average, allocated 26.96%, 33.69%, and 37.35% of the joint production payoff to subject B when only subject A was right. In the corresponding situations A_rB_w in the UGT (UGT1R), proposers still offered 39.26% (34.53%) of the joint output to the responders in their groups. Interestingly, this rather generous allocation occurred despite the fact that proposers had superior bargaining power *and* had created the joint production payoff available for distribution. For all treatments, offers in the asymmetric cases A_rB_w were significantly lower than 50% (t-test, level of significance at 1%).

When, in contrast, only subject B was right, third parties in the TPT, TPTSQM, and TPTSQNM allocated, on average, 71.07%, 66.43%, and 60.52% to subjects B, and in the UGT (UGT1R), proposers offered responders 61.52% (60.99%) of the output. For all treatments, offers in situation A_wB_r were significantly larger than 50% (t-test, level of significance at 1%).

Figure **2** and Figure **3** illustrate *Result 2* graphically. Figure **2** depicts the frequencies of shares distributed to subjects B (responders) if only subject A (the proposer) gave the correct answer (averaged over all nine matrices). In all treatments, most subjects allocated 50% or less of the joint production payoff to subject B.

Insert Figure 2 about here

Figure **3** shows the average frequencies of shares distributed to subjects B when only subject B was right. It is clearly evident that in situation A_wB_r, most third parties and most proposers allocated at least half of the output to subject B.

Insert Figure 3 about here

On the whole, the results for the asymmetric cases are consistent with fairness based on accountability for giving the correct answer.

Our findings thus stand in contrast to the results of numerous experiments on Ultimatum Game bargaining on *exogenously* given endowments; earlier studies have shown that the proposers' modal (mean) offer is typically 50% (averages between 40% and 50%) of the distributable amount (see *Camerer and Thaler* (1995), *Roth* (1995), *Oosterbeek et al.* (2004)). Thus, the fact that the subjects in our design were distributing an endogenous joint output clearly affected their distributional choices. However, allocations to subjects B within a treatment condition were heterogeneous. In the UGT, for example, 26.9% of the proposers generously offered the responder a share of 50% of the joint production payoff, even when only the proposer knew the correct answer (see Figure **2**). When only the responder was right, the frequency of equal offers in the UGT decreased to 20.4% (see Figure **3**); most proposers suggested distributing two-thirds of the joint production payoff to the responder. Thus, in the asymmetric cases, we find heterogeneity in the fairness norm applied, either because the accountability principle was not equally important for all of the proposers (*Brandstätter and Königstein* (2001)), or because the proposers faced uncertainty about the responders' fairness norms. Equal sharing might thus be seen as a reference point.

V.1.c Result 3

In the symmetric case in which both subjects were right (A_rB_r), the joint production payoff of 100 was shared equally in each matrix of the treatments with simultaneous framing (TPT, UGT, and UGT1R).

Table 6 shows that in the symmetric case in which both subjects knew the correct answer, subjects, on average, again shared the resulting 100 Points equally. When both subjects knew the correct answer, third parties in the TPT, TPTSQM, and TPTSQNM distributed approximately 50%, 51%, and 51% of the 100 Points to subject B, and proposers in the UGT (UGT1R) offered responders approximately 49% (48%) of the joint output.

Figure **4** illustrates the result that almost all third parties and most of the proposers offered half of the joint production payoff to subject B.

Insert Figure 4 about here

When we compare shares distributed to subjects B for the symmetric case A_rB_r across treatment conditions, we find no differences between the UGT, the UGT1R, and the TPT at conventional levels of significance (two-sided t-test or Wilcoxon signed-rank test). When we compare the TPTSQM or the TPTSQNM to the other treatments, however, we find significant differences for some matrices in the allocation of the joint output of A_rB_r between these treatments and the others, and there are significant differences between the TPTSQM and the TPTSQNM for matrices JP_1, JP_2, JP_7, JP_8, and JP_9 (two-sided t-test). Using data from the symmetric case $A_r B_r$, a regression of the shares allocated to subject B on dummy variables for the different payoff matrices shows that for the UGT, the UGT1R, and the TPT, the payoff matrix did not affect the distribution of the 100 Points resulting from both subjects answering the question correctly. For these treatments with simultaneous framing, *Result 3* is thus robust to the payoff matrix. Similar regressions for the treatments with sequential framing, however, show that matrices JP_2 and JP_3 in the TPTSQNM and matrices JP_4 to JP_9 in the TPTSQM produced results that were significantly different from those of the other matrices.

Libertarian fairness also predicts shares of 0.5 for all matrices if participants focus on losses, and for matrices JP_1, JP_5, and JP_9 if subjects focus on gains. For the remaining matrices (for which libertarian fairness predicts shares other than 0.5, as indicated by the variables *Shapley_Gain* defined in Table 4 and listed in Table 7), mean offers in the TPT, UGT, and UGT1R treatments were also (approximately) at 50%. The participants in these treatments did not seem to take different individual productivities into account, but rather shared the joint production payoff equally. As a result, for situation $A_r B_r$, our data for the TPT, UGT, and UGT1R are not in line with the predictions of libertarian fairness, but rather with equal sharing and with the accountability principle.

V.1.d Result 4

Libertarian fairness only played a major role in the TPTSQM – not in the UGT, the UGT1R, the TPT, or the TPTSQNM.

Results 1 and *3* for the symmetric cases state that data from the TPT, UGT, and UGT1R are consistent with both equal sharing (strict egalitarianism) and the accountability principle (or liberal egalitarianism), and *Result 2* for the asymmetric cases indicates that the accountability principle is more appropriate to explain our data. For the treatments with simultaneous framing, we do not find evidence that subjects apportioned Points based on individual marginal productivities (libertarianism), since participants did not consistently distribute in proportion to marginal gains or marginal losses.

If subjects indeed viewed a distribution according to individual marginal productivities as fair and focused on gains, we should observe shares distributed to subject B, α_m^B, in the following order for our payoff matrices: $\alpha_{1,4,7}^B < \alpha_{2,5,8}^B < \alpha_{3,6,9}^B$ for situation $A_w B_r$, and $\alpha_{1,2,3}^B > \alpha_{4,5,6}^B > \alpha_{7,8,9}^B$

for situation $A_r B_w$. If our subjects focused on losses, we should find the opposite ordering of shares allocated to subject B, i.e., $\alpha^B_{1,4,7} > \alpha^B_{2,5,8} > \alpha^B_{3,6,9}$ for situation $A_w B_r$, and $\alpha^B_{1,2,3} < \alpha^B_{4,5,6} < \alpha^B_{7,8,9}$ for situation $A_r B_w$. A comparison of average offers across the respective three matrices results in 0.73 > 0.68 < 0.71 (TPT), 0.62 < 0.63 > 0.60 (UGT), and 0.62 > 0.60 < 0.61 (UGT1R) for $A_w B_r$, and in 0.28 < 0.29 > 0.24 (TPT), 0.39 < 0.40 > 0.39 (UGT), and 0.35 > 0.34 < 0.35 (UGT1R) for $A_r B_w$. Evidently, shares are similar across these treatments; even if we ignore significance, we do not see a consistent order in our data. We therefore conclude that libertarian fairness does not explain our data for the asymmetric cases of the TPT, the UGT, and the UGT1R, irrespective of whether subjects focused on gains or on losses.

In the treatments with simultaneous framing, we presented nine different matrices JP_m to our subjects, i.e., 9 x 4 joint production payoffs. This raises the question of whether marginal productivities are not relevant because they are *exogenously* given (i.e., subjects do not have an influence on the magnitude of the output increase resulting from knowing the correct answer), or because subjects *do not notice* differences in marginal productivities across different payoff matrices. To shed more light on this issue, we conducted the TPTSQM, which emphasized the sequential framing of individual marginal productivities. In situation $A_r B_r$, subject A appeared to be the first mover in answering questions. In the TPTSQNM, we used the same sequential framing; however, in contrast to the TPTSQM, we informed subjects about the resulting joint production payoff (instead of the marginal productivities).

If we again calculate the average shares allocated to subject B for the respective matrices, we find that the ordering predicted for subjects adhering to ideals of libertarian fairness and focusing on gains (namely, $\alpha^B_{1,4,7} < \alpha^B_{2,5,8} < \alpha^B_{3,6,9}$ in situation $A_w B_r$ and $\alpha^B_{1,2,3} > \alpha^B_{4,5,6} > \alpha^B_{7,8,9}$ in situation $A_r B_w$) is indeed present in the data from our TPTSQM (0.63 < 0.67 < 0.68 ($A_w B_r$) and 0.35 > 0.33 ≥ 0. 33 ($A_r B_w$)), but not in the data from our TPTSQNM (0.61 > 0.6 < 0.61 ($A_w B_r$) and 0.38 = 0.38 > 0.37 ($A_r B_w$)). Moreover, using data from the symmetric case $A_r B_r$ for a regression of the shares allocated to subject B on dummy variables for the different payoff matrices, we find that in matrices JP_7, JP_8, and JP_9, in which A's marginal productivity is 70 Points, the shares allocated to B are significantly lower than 50% in the TPTSQM (but not in the TPTSQNM). In matrices JP_1, JP_2, and JP_3, in which A's marginal productivity at 10 Points is comparatively low, the shares allocated to B are significantly higher than 50% in the

TPTSQM (in the TPTSQNM, shares are significantly higher than 50% only for JP_3, and are actually lower for JP_2). These findings indicate that an emphasis on marginal productivities shifts distributional choices to the libertarian fairness ideal, but that the framing as a sequential production process does not. Marginal productivities thus affect distributional choices only when they are readily apparent.

V.2 Regression Analysis

V.2.a Additional Evidence for Results 1-4

In the following section, we report the results of regression analyses conducted in order to confirm our results from the descriptive statistics and to determine the relative importance of the different fairness norms. Table 8 provides the results of separate regressions for each treatment, with shares allocated to subjects B as dependent variables and with subjects as clusters. To investigate whether group members who had answered their question correctly received a larger share than those who were wrong (*Result 2*), we include the explanatory variables *A_right* (*B_right*) as dummy variables. *A_right* (*B_right*) takes on the value of 1 when subject A (subject B) knew the right answer to the question, and zero otherwise. To test whether subjects distributed according to the libertarian fairness ideal (*Result 4*), we include the variables *Shapley_Gain_-0.5* (*Shapley_Loss_-0.5*), which are shares predicted for subjects B should participants adhere to the libertarian ideal and focus on gains (on losses), as indicated in Table 4 (Table 5), minus 0.5.

Insert Table 8 about here

In all of the regressions, the coefficient of *A_right* (*B_right*) is negative (positive) and highly significant. The signs of these coefficients support *Result 2*, which states that the subject who knew the correct answer will be offered a higher share of the joint production payoff than the group member who was wrong. When subject A knew the correct answer, subjects allocated a lower share to subject B. When, in contrast, subject B was right, third parties and proposers rewarded subject B with a higher share. When both group members were right, the coefficients of *A_right* and *B_right* roughly sum up to zero in each of these regressions (*Result 3*).

Taking the respective constants into account, the joint production payoff is shared almost equally in all of our treatments when both group members were equally responsible for its realization (*Result 3*) and when both group members were wrong (*Result 1*). For our treatments without an emphasis on marginal productivities (TPT, TPTSQNM, UGT, and UGT1R), *Shap-*

ley_Gain_-0.5 is not significant; *Shapley_Loss_-0.5* is borderline significant only in the TPT. In the regression for the TPTSQM, in contrast, the coefficient of *Shapley_Gain_-0.5* is highly significant. In line with *Result 4*, individual marginal productivities as reflected in the Shapley Values thus did not affect distributional choices in the treatments in which we did not make marginal productivities highly apparent, but considerably shaped the allocation when the realization of marginal productivities was framed as a sequential process. The latter result is contrary to the accountability principle stated by *Konow* (2000) and *Konow* (2003), since marginal productivities were behaviorally relevant in our TPTSQM even though subjects had no influence on this variable. The behavioral effect of the framing that we implemented in the TPTSQM might result in a shift in the individual perception of responsibility, emphasizing the amount of additional Points earned by knowing the correct answer rather than the knowledge itself.

Because we framed the situation in the TPTSQM and in the TPTSQNM in such a way that subject A appeared to be the first mover, our variable *Shaply_Gain_-0.5* might be inappropriate. Consequently, we calculated the variable *Shaply_Gain_SQ_-0.5*, considering A as the first mover. Regressions with *Shaply_Gain_SQ_-0.5* rather than *Shaply_Gain_-0.5* and *Shaply_Loss_-0.5* yield results that are similar to those reported in Table 8.[4] Thus, the shares that third parties in the TPTSQM distributed to subjects B increased with B's marginal productivity, whereas shares in the TPTSQNM were insensitive to marginal productivities.

To confirm the existence of this shift in the fairness norm, we ran a separate regression for each subject, with the shares allocated to subject B as the dependent variable and *Shapley_Gain*, *Shapley_Loss*, *A_right*, and *B_right* as explanatory variables. Based on the coefficients from these 152 regressions, we can cluster our subjects into three different fairness types: Egalitarians have coefficients of approximately 0. Libertarians are characterized by the lowest constants and the highest (positive) coefficients for *Shapley_Gain_-0.5*. In addition to rather high constants, participants who allocated according to the accountability principle have large (negative) coefficients for *A_right* and large (positive) coefficients for *B_right*. For each treatment, Table 9 shows the percentage of subjects that can be classified as one of these types. With 42%, the TPTSQM was the treatment with the lowest percentage of subjects distributing Points according to the accountability principle. In addition, the percentage of Libertarians at 35% was higher in the TPTSQM than in any other treatment; in other treatments, the percentage of Libertarians ranged from 7% to 8%. Our results for the remaining treatments

[4] Since the correlation between *Shaply_Gain_SQ_-0.5* and *A_right* is stronger than that between *Shaply_Gain_-0.5* and *A_right*, the new regression has an even lower explanatory power.

are qualitatively in line with *Cappelen et al.* (2007), *Cappelen et al.* (2011), and *Cappelen et al.* (2010).

Insert Table 9 about here

In summary, we find that the accountability principle can explain our data for the treatments with sequential framing (i.e., TPT, UGT, and UGT1R); libertarianism does not seem to be relevant. In these treatments, we showed subjects all nine matrices (i.e., we informed them about their marginal productivities), and the differences between the payoff matrices were also observable. Nonetheless, distributional choices in these treatments were insensitive to individual marginal productivities. In the TPTSQM, in contrast, we implemented an economically completely equivalent situation, but we made marginal productivities more apparent by accentuating the sequential framing of the production process. This framing had significant behavioral effects: The prevalent fairness norm shifted to libertarianism. The reason for this shift was determined to be the emphasis on marginal productivities rather than the sequential framing, since in the TPTSQNM, libertarianism was not the prevalent fairness norm.

V.2.b Result 5

Subjects in the UGT and in the UGT1R principally adhered to the same fairness norm as subjects in the TPT, but subjects in the UGT and in the UGT1R were "less extreme" in their distributional choices than subjects in the TPT.

A comparison of shares in the asymmetric cases across treatments TPT, UGT, and UGT1R reveals that these allocations are qualitatively similar (see Table 7). In addition, the coefficients in the regressions for the TPT, the UGT, and the UGT1R are similar (see Table 8). Thus, proposers in the UGT and in the UGT1R, in principle, seemed to follow the same fairness norm as third parties in the TPT. The predominance of the accountability principle was therefore not affected when implicated parties with asymmetric bargaining power distributed a joint output.

In the asymmetric cases, however, subjects in the Ultimatum Game Treatments were less "extreme" than subjects in the TPT, i.e., proposers rewarded the subject who knew the answer less than third parties did. If we compare allocations to subjects B between the UGT and UGT1R on the one hand and the TPT on the other (see Figure 2 and Figure 3), we see fewer large offers and fewer meager offers in the UGT / UGT1R than in the TPT. In Section IV, we predicted that any excessively large allocations to subjects B observed in the TPT should disappear in the UGT / UGT1R, because overly generous offers would decrease proposers' payoffs. Likewise, relatively small allocations to subjects B should be less frequent in the UGT / UGT1R than in the TPT, because proposers would fear responders' rejections. This reasoning is confirmed using regressions containing interaction terms: The coefficients of *A_right* and *B_right* in the UGT and the TPT are statistically different (1% level of significance), as are those in the UGT1R and the TPT (5% level of significance).[5]

Another explanation for this result could be that fairness norms shift towards being more egalitarian when the relationship becomes more personal (*Konow et al.* (2009)). *Result 5*, however, stands in contrast to the findings of *Konow* (2000), *Croson and Konow* (2009), and *Konow et al.* (2009), which report that the decisions of quasi-spectators are significantly *less* widely scattered than those of stakeholders.

To rule out proposers' *self-serving bias* (i.e., subjects interpret the relative importance of different fairness norms in their own best interest) as an explanation for *Result 5*, we compare the results of the TPT to those of the UGT / UGT1R. The significant difference between the coefficients of *B_right* in the UGT / UGT1R and the TPT mentioned above is consistent with self-serving bias, since biased proposers should place more weight on equal sharing when only the responder knows the right answer. However, the absolute value of the coefficient of *A_right* is significantly *smaller* for the UGT / UGT1R than for the TPT at a 1% (5%) level of significance, i.e., proposers did not overemphasize the relevance of accountability when only they knew the correct answer. In line with *Cappelen et al.* (2011) and *Konow et al.* (2009) (but in contrast to *Cappelen et al.* (2007) and *Babcock and Loewenstein* (1997)), we thus do not find consistent evidence for proposers' self-serving bias.

A second test addresses the *behavioral effects of the role reversal* we implemented in the UGT as an explanation for the observed difference between the UGT and the TPT. In particular, the role reversal could have resulted in a less pronounced role perception in the UGT and thus in offers centered on 50%. From Table 6 and Table 7, however, we see that there are no qualitative differences between the UGT and the UGT1R, and a t-test yields no statistically significant differences. Table 8 shows that the coefficients for the UGT1R are similar to those for the UGT, and regressions with interaction terms for treatments also show no significant differences between the UGT and the UGT1R. Moreover, the coefficients of *A_right* and *B_right* in the UGT1R regression are smaller than those in the TPT regression and are statistically significant at the 5%-level. Consequently, we conclude that a less pronounced role per-

[5] Regressions on data sets of two treatments including interaction terms of our explanatory variables with treatment dummies show that neither the coefficients of *A_right* and *B_right* for the UGT and the UGT1R nor the constants are statistically different.

ception due to the role reversal in the UGT is not the primary reason for the decrease in the importance of fairness based on accountability from the TPT to the UGT.

In principle, the role reversal in the UGT could also have encouraged subjects to justify their decisions by *hindsight*: If participants in subgroup I (first proposer, then responder) subconsciously validated the decisions they made as proposers when setting their acceptance thresholds in the second half of the session, these participants as responders would set lower minimum thresholds than responders in subgroup II (first responder, then proposer) did in the first half of the session. Likewise, if participants in subgroup II sought to justify their decisions as responders when making offers in the second half of the session, these participants would make more generous offers than proposers in subgroup I. However, when we compare offers and minimum thresholds between subgroups I and II by conducting a t-test, we find that offers in only 2 out of 36 situations are statistically different at the 5%-level, and that none of the minimum claims is statistically different. Thus, we can conclude that this kind of decision bias is not present in our data and does not explain *Result 5*.

V.2.c Result 6

In the UGT and in the UGT1R, responders principally adhered to the same fairness norm as proposers.

In the symmetric cases of the UGT (UGT1R), responders claimed, on average, a share of 35.69% (40.20%) when both subjects gave the wrong answer and 33.25% (37.63%) when both group members answered correctly (see Table 6). In the asymmetric cases, however, responders' acceptance thresholds were sensitive to the correctness of the group members' answers: When only the proposer gave the right answer, responders in the UGT (UGT1R) claimed, on average, 23.30% (26.96%) of the joint output, i.e., a lower share than in the symmetric cases. But when only the responder was right, the average acceptance thresholds in the UGT (UGT1R) increased to 40.88% (47.36%), i.e., above the share claimed in the symmetric cases of the respective treatment. Thus, we conclude that proposers' minimum claims are also in line with the accountability principle.

Figure **5** and Figure **6** depict the frequencies of responders' acceptance thresholds for the symmetric and asymmetric cases of the UGT and the UGT1R, respectively.

Insert Figure 5 about here

Insert Figure 6 about here

We can see that in the symmetric cases, some responders claimed 50% of the joint output. However, in the asymmetric cases, and in particular in situation $A_w B_r$ (where it is less obvious what allocations will be viewed as "fair"), we observe a considerable heterogeneity in minimum claims. Although some responders adhered to equal sharing, there were also responders who seemed to adhere to the accountability principle.

To determine whether responders operated according to the same fairness norm as proposers, we ran regressions for the UGT and the UGT1R with the minimum shares that responders claimed as the dependent variable and with matching groups as clusters. Table 10 shows that the coefficients of the variables *Shapley_Gain_-0.5* and *Shapley_Loss_-0.5* are again not significant. Thus, marginal productivities do not determine minimum claims. The coefficient of *A_right* (*B_right*) is again negative (positive) and is highly significant. Although the absolute values of the coefficients in the regression for responders' claims are lower than those in the regression for proposers' offers, the difference in *A_right* is not statistically significant, and the difference in *B_right* is only borderline significant. Since proposers' offers and responders' claims depend similarly on the variables included in the respective regressions, we conclude that proposers and responders principally shared the same fairness norm.

Insert Table 10 about here

VI. Concluding Remarks

This paper presents the results of an experimental study designed to investigate the prevalence of different norms of distributive fairness in team production processes in which individual contributions are not readily observable. In our experiment, each of two subjects within a team had to answer a knowledge question to jointly generate an output. The joint production payoff was 20 Points when both subjects gave the wrong answer and 100 Points when both subjects answered the question correctly. Payoffs in the asymmetric cases in which only one subject was right varied across the nine payoff matrices we used. We included matrices with synergies (i.e., the additional payoff increases of both team members were less than 80 Points) and negative externalities (i.e., the sum of the marginal productivities exceeded 80 Points). Our design allowed us to investigate whether subjects (a) shared the joint production payoff equally (egalitarianism), (b) rewarded subjects who knew the correct answer (accountability principle), or (c) took into account (exogenously given) individual marginal productivities (libertarianism).

We find that fairness based on accountability was the most commonly applied norm in a Third Party Treatment (TPT) in which a person outside the team distributed the joint production payoff. Thus, people are rewarded only for giving the right answer, not for the resulting payoff increase. Our results indicate that the predominance of the accountability principle was unaffected when subjects allocated an output they had previously generated within their team, using the rules of the Ultimatum Game (UGT). This means that membership in the team that realizes the joint output does not seem to affect the importance of the accountability principle. Moreover, both proposers' offers and responders' minimum claims were in line with the accountability principle. Since the payoffs of the subjects in the UGT depended on their distributive choices, however, they did not reward the team member who knew the correct answer to the same extent that subjects in the TPT did.

In these treatments, we implemented a simultaneous framing, in the sense that we showed participants the resulting payoff matrices. One reason for our finding that subjects did not take marginal productivities into account might be that the reference point for the distribution of the joint output was not clear, since we allowed for synergies and negative externalities. Another explanation is that the payoff increase was exogenously given (i.e., not under our subjects' control), and therefore subjects regarded the consideration of marginal productivities as unfair. To shed more light on this issue, we conducted a Third Party Treatment with a sequential framing and an emphasis on marginal productivities (TPTSQM): Subjects started with 20 Points, and we explicitly informed them about the payoff increase that would result if one or both subjects gave the correct answer. In addition, we conducted another Third Party Treatment with the same sequential framing of the production process, but in which we informed participants about the resulting joint production payoff instead of their individual marginal productivities (TPTSQNM). Our findings indicate that the framing presented in the TPTSQM caused a shift from fairness based on accountability to libertarianism, whereas subjects in the TPTSQNM still rewarded endogenous inputs (accountability). Thus, subjects consider marginal productivities only when they are truly obvious.

Based on our results, marginal productivities should be integrated into performance measures only if they are highly apparent, as is the case in piece-rate compensation schemes. In cases with unobservable individual marginal productivities, the subjective performance measures that are frequently used as proxies for these marginal productivities are often viewed as unfair. Our findings offer an explanation of why management compensation is often regarded as unfair: If workers spend the same (or even more) work time in the company as the managers do and exert the same effort level, the high levels of compensation paid to managers can be seen as excessive and unfair, in particular when the increase in firm value achieved by the management is not readily observable. In such cases, many people would feel that the compensation should only take into account factors that are under an individual's control (e.g., input, effort levels, investment costs, etc.) or depend on observable output increases ("pay for performance").

Our experimental design is not without limitations. First, inputs were in the form of answering multiple-choice knowledge questions. Thus, individual productivities were not observable, and we cannot differentiate between effort, talent, and luck. As a consequence, we are unable to disentangle the accountability principle from liberal egalitarianism. Future experiments could employ a design that controls for talent and / or intention, for example, by letting subjects determine the level of difficulty of the questions they have to answer. Second, the relative importance of fairness norms might be different in cases in which people lose rather than gain money. Thus, our findings might not extend to decisions on tax payments or donations to Third World Countries. Further research thus could address the prevalence of different fairness norms in situations in which people spent part of their wealth.

Tables and Figures

Subject B (columns) / Subject A (rows)

1	w	r
w	20	30
r	30	100

2	w	r
w	20	60
r	30	100

3	w	r
w	20	90
r	30	100

4	w	r
w	20	30
r	60	100

5	w	r
w	20	60
r	60	100

6	w	r
w	20	90
r	60	100

7	w	r
w	20	30
r	90	100

8	w	r
w	20	60
r	90	100

9	w	r
w	20	90
r	90	100

Table 1: Joint production payoff matrices (payoffs in Points)

	Number of subjects	Average payments to participants
TPT	24	21.70 € / 34.29 $
UGT	24	20.00 € / 31.60 $
UGT1R	50	14.80 € / 20.78 $
TPTSQM	26	17.40 € / 24.36 $
TPTSQNM	28	13.63 € / 18,38 $
Total	***152***	---

Table 2: Number of subjects and average payments to participants, by treatment condition

		A_wB_r		
		30 Points	**60 Points**	**90 Points**
A_rB_w	**30 Points**	40; 40	25; 55	10; 70
	60 Points	55; 25	40; 40	25; 55
	90 Points	70; 10	55; 25	40; 40

Table 3: Shapley Values for subjects A and B (in Points), by joint production payoff matrix

Subject A \ Subject B	1		2		3	
	w	r	w	r	w	r
w	0.50	0.67	0.50	0.83	0.50	0.89
r	0.33	0.50	0.33	0.65	0.33	0.80
	4		**5**		**6**	
	w	r	w	r	w	r
w	0.50	0.67	0.50	0.83	0.50	0.89
r	0.17	0.35	0.17	0.50	0.17	0.65
	7		**8**		**9**	
	w	r	w	r	w	r
w	0.50	0.67	0.50	0.83	0.50	0.89
r	0.11	0.20	0.11	0.35	0.11	0.50

Table 4: Predictions regarding shares distributed to B
(fairness based on marginal productivities with a focus on gains)

Game	Subject A \ Subject B	w	r
1	w	0.5	1.0
1	r	0	0.50
2	w	0	0.83
2	r	0	0.50
3	w	0	0.56
3	r	0	0.50
4	w	1.0	1.0
4	r	0.17	0.50
5	w	0.50	0.83
5	r	0.17	0.50
6	w	0	0.56
6	r	0.17	0.50
7	w	1.0	1.0
7	r	0.44	0.50
8	w	1.0	0.83
8	r	0.44	0.50
9	w	0.5	0.56
9	r	0.44	0.50

Table 5: Predictions regarding shares distributed to B
(fairness based on marginal productivities with a focus on losses)

		Allocations to subject B		Minimum claims of responders	
		B was wrong	B was right	B was wrong	B was right
TPT	A was wrong	0.5016 (0.1139)	0.7107 (0.2293)	-	-
	A was right	0.2696 (0.2172)	0.5017 (0.1015)	-	-
TPTSQM	A was wrong	0.4955 (0.0768)	0.6643 (0.1423)	-	-
	A was right	0.3369 (0.1399)	0.5058 (0.1109)	-	-
TPTSQNM	A was wrong	0.5056 (0.1122)	0.6052 (0.1718)	-	-
	A was right	0.3735 (0.1676)	0.5050 (0.1070)	-	-
UGT	A was wrong	0.4896 (0.0652)	0.6152 (0.1054)	0.3569 (0.1596)	0.4088 (0.2169)
	A was right	0.3926 (0.0901)	0.4875 (0.0527)	0.2330 (0.1481)	0.3325 (0.1761)
UGT1R	A was wrong	0.4818 (0.0843)	0.6099 (0.1303)	0.4020 (0.1898)	0.4736 (0.2363)
	A was right	0.3453 (0.1157)	0.4819 (0.0691)	0.2696 (0.1750)	0.3763 (0.1812)

Table 6: Average shares distributed to subject B, by treatment condition, and minimum claims of the responders in the UGT and UGT1R
(averages across all nine matrices; standard deviations in parentheses)

	JP$_m$ = (20; $\boldsymbol{A_wB_r}$; $\boldsymbol{A_rB_w}$; 100)								
	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
	30; 30	**60; 30**	**90; 30**	**30; 60**	**60; 60**	**90; 60**	**30; 90**	**60; 90**	**90; 90**
Shapley_Gain									
A_wB_w	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
A_wB_r	0.67	0.83	0.89	0.67	0.83	0.89	0.67	0.83	0.89
A_rB_w	0.33	0.33	0.33	0.17	0.17	0.17	0.11	0.11	0.11
A_rB_r	0.50	0.65	0.80	0.35	0.50	0.65	0.20	0.35	0.50
Shapley_Loss									
A_wB_w	0.50	0.00	0.00	1.00	0.50	0.00	1.00	1.00	0.50
A_wB_r	1.00	0.83	0.56	1.00	0.83	0.56	1.00	0.83	0.56
A_rB_w	0.00	0.00	0.00	0.17	0.17	0.17	0.44	0.44	0.44
A_rB_r	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
TPT Shares for B									
A_wB_w	0.48	0.57	0.49	0.50	0.52	0.50	0.50	0.48	0.48
A_wB_r	0.70	0.67	0.67	0.73	0.70	0.77	0.78	0.67	0.70
A_rB_w	0.31	0.25	0.27	0.30	0.32	0.25	0.28	0.23	0.23
A_rB_r	0.52	0.49	0.50	0.51	0.53	0.49	0.50	0.50	0.48
UGT Proposers' Offers									
A_wB_w	0.48	0.49	0.49	0.50	0.50	0.49	0.50	0.48	0.48
A_wB_r	0.62	0.64	0.59	0.60	0.62	0.61	0.63	0.62	0.61
A_rB_w	0.38	0.38	0.41	0.40	0.39	0.39	0.40	0.40	0.38
A_rB_r	0.49	0.49	0.49	0.47	0.49	0.49	0.49	0.49	0.49
UGT Responders' Claims									
A_wB_w	0.37	0.36	0.36	0.36	0.35	0.36	0.35	0.34	0.36
A_wB_r	0.41	0.44	042	0.40	0.43	0.40	0.42	0.38	0.38
A_rB_w	0.24	0.22	0.26	0.24	0.23	0.24	0.22	0.23	0.22
A_rB_r	0.33	0.33	0.34	0.33	0.33	0.34	0.33	0.33	0.33
UGT1R Proposers' Offers									
A_wB_w	0.48	0.47	0.50	0.47	0.50	0.48	0.47	0.49	0.47
A_wB_r	0.62	0.59	0.60	0.62	0.60	0.59	0.60	0.61	0.64
A_rB_w	0.35	0.36	0.34	0.33	0.33	0.35	0.36	0.34	0.34
A_rB_r	0.50	0.47	0.47	0.48	0.49	0.49	0.48	0.48	0.46
UGT1R Responders' Claims									
A_wB_w	0.38	0.41	0.40	0.42	0.40	0.42	0.40	0.39	0.39
A_wB_r	0.46	0.49	0.49	0.50	0.46	0.47	0.48	0.45	0.46
A_rB_w	0.27	0.28	0.27	0.27	0.26	0.28	0.27	0.27	0.27
A_rB_r	0.37	0.38	0.38	0.37	0.37	0.38	0.38	0.38	0.37
TPTSQM Shares for B									
A_wB_w	0.49	0.50	0.49	0.52	0.50	0.51	0.48	0.46	0.51
A_wB_r	0.64	0.68	0.68	0.61	0.67	0.67	0.64	0.67	0.71
A_rB_w	0.36	0.34	0.36	0.34	0.32	0.33	0.34	0.33	0.31
A_rB_r	0.58	0.55	0.58	0.50	0.50	0.51	0.45	0.44	0.45
TPTSQNM Shares for B									
A_wB_w	0.47	0.52	0.48	0.56	0.53	0.48	0.51	0.51	0.50
A_wB_r	0.60	0.61	0.61	0.55	0.61	0.63	0.65	0.62	0.57
A_rB_w	0.36	0.37	0.34	0.38	0.41	0.38	0.39	0.36	0.39
A_rB_r	0.51	054	0.51	0.45	0.50	0.51	0.54	0.49	0.50

Table 7: Summary statistics for the TPT, the TPTSQM, and the TPTSQNM conditions (shares that third parties allocated to subjects B; averages by payoff matrix) and the UGT and UGT1R conditions (shares that proposers offered and acceptance thresholds that responders set; averages by payoff matrix)

Variables	TPT	TPTSQM	TPTSQNM	UGT	UGT1R
Shapley_Gain_-0.5	-0.0396 (0.0383)	0.201*** (0.0609)	-0.0319 (0.0586)	0.000709 (0.0234)	-0.0141 (0.0149)
Shapley_Loss_-0.5	-0.0324* (0.0163)	-0.00475 (0.00941)	0.0343 (0.0211)	0.0107 (0.00710)	-0.00624 (0.00683)
A_right	-0.242*** (0.0399)	-0.100*** (0.0208)	-0.136*** (0.0231)	-0.109*** (0.0162)	-0.138*** (0.0189)
B_right	0.242*** (0.0367)	0.111*** (0.0200)	0.135*** (0.0238)	0.107*** (0.0154)	0.138*** (0.0195)
Constant	-0.00428 (0.0073)	-0.00451** (0.00216)	-0.0024 (0.00526)	-0.00273 (0.00760)	-0.0203* (0.00117)
# Observations	864	936	1008	864	900
R^2	0.443	0.499	0.25	0.485	0.454
# Clusters	24	26	28	24	25

* p < 0.1; ** p < 0.05; *** p < 0.01

Table 8: Regressions with the shares offered to subjects B / responders (minus 0.5) as the dependent variable (robust standard errors in parentheses)

	Egalitarianism	Libertarianism	Accountability
TPT	13%	8%	79%
TPTSQM	23%	35%	42%
TPTSQNM	32%	7%	61%
UGT	33%	8%	58%
UGT1R	32%	0%	68%

Table 9: Percentage of different fairness types in our treatment conditions

Variables	**UGT**	**UGT1R**
Shapley_Gain_-0.5	0.00431 (0.0159)	-0.000189 (0.00954)
Shapley_Loss_-0.5	-0.00708 (0.00918)	-0.00553 (0.00756)
A_right	-0.101*** (0.0181)	-0.117*** (0.0170)
B_right	0.0765*** (0.0199)	0.0909*** (0.0180)
Constant	-0.155*** (0.0309)	-0.107*** (0.0367)
# Observations	864	900
R^2	0.112	0.120
# Clusters	24	25

* p < 0.1; ** p < 0.05; *** p < 0.01

Table 10: Regressions with the minimum shares subjects B / responders claimed for themselves as the dependent variable (robust standard errors in parentheses)

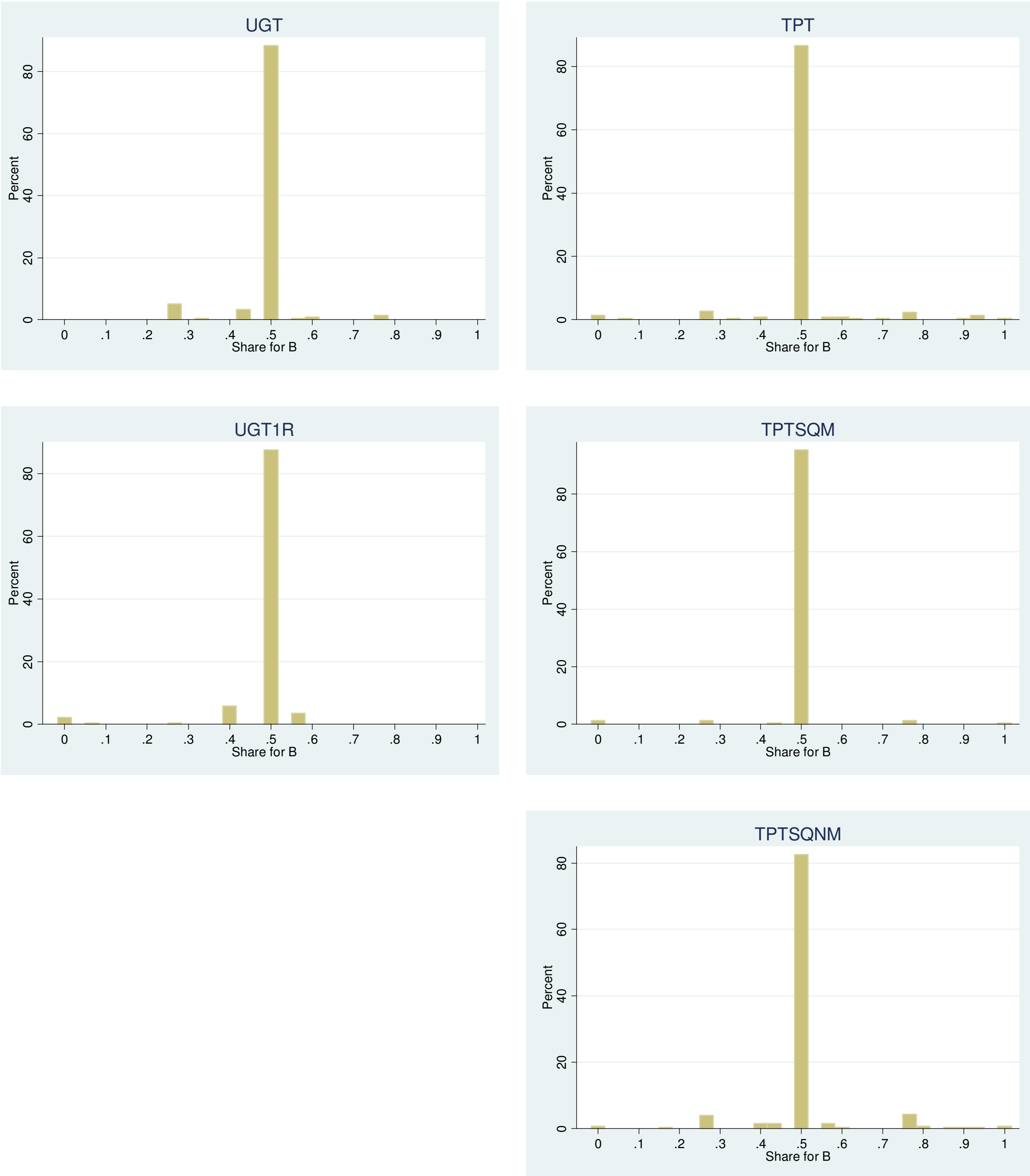


Figure 1: Frequencies (in percent) of shares that third parties and proposers distributed to subjects B / responders when both subjects gave the wrong answer (average frequencies for all nine matrices)

Figure 2: Frequencies (in percent) of shares that third parties and proposers distributed to subjects B / responders when only subject A / the proposer was right (average frequencies for all nine matrices)

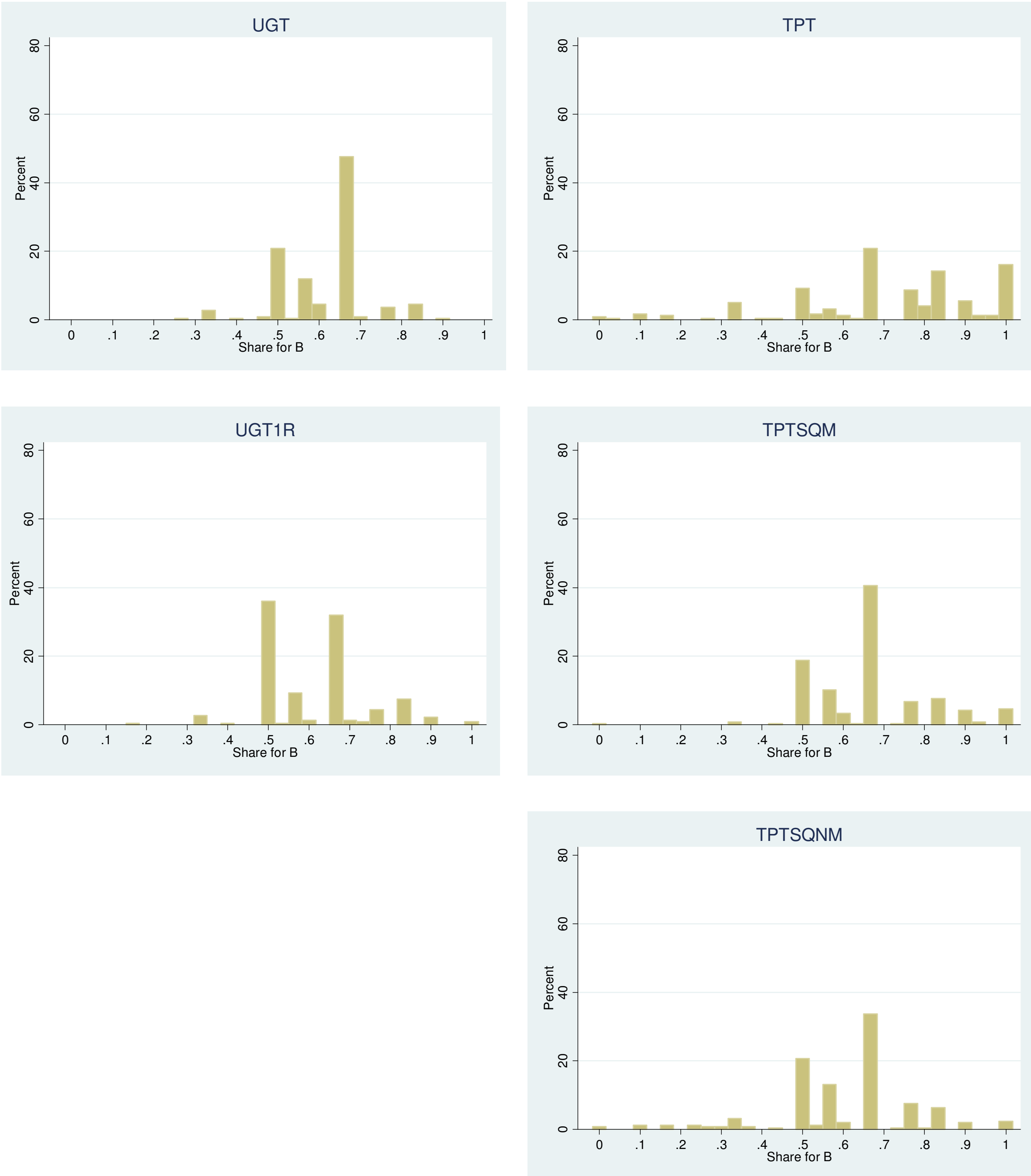


Figure 3: Frequencies (in percent) of shares that third parties and proposers distributed to subjects B / responders when only subject B / the responder was right (average frequencies for all nine matrices)

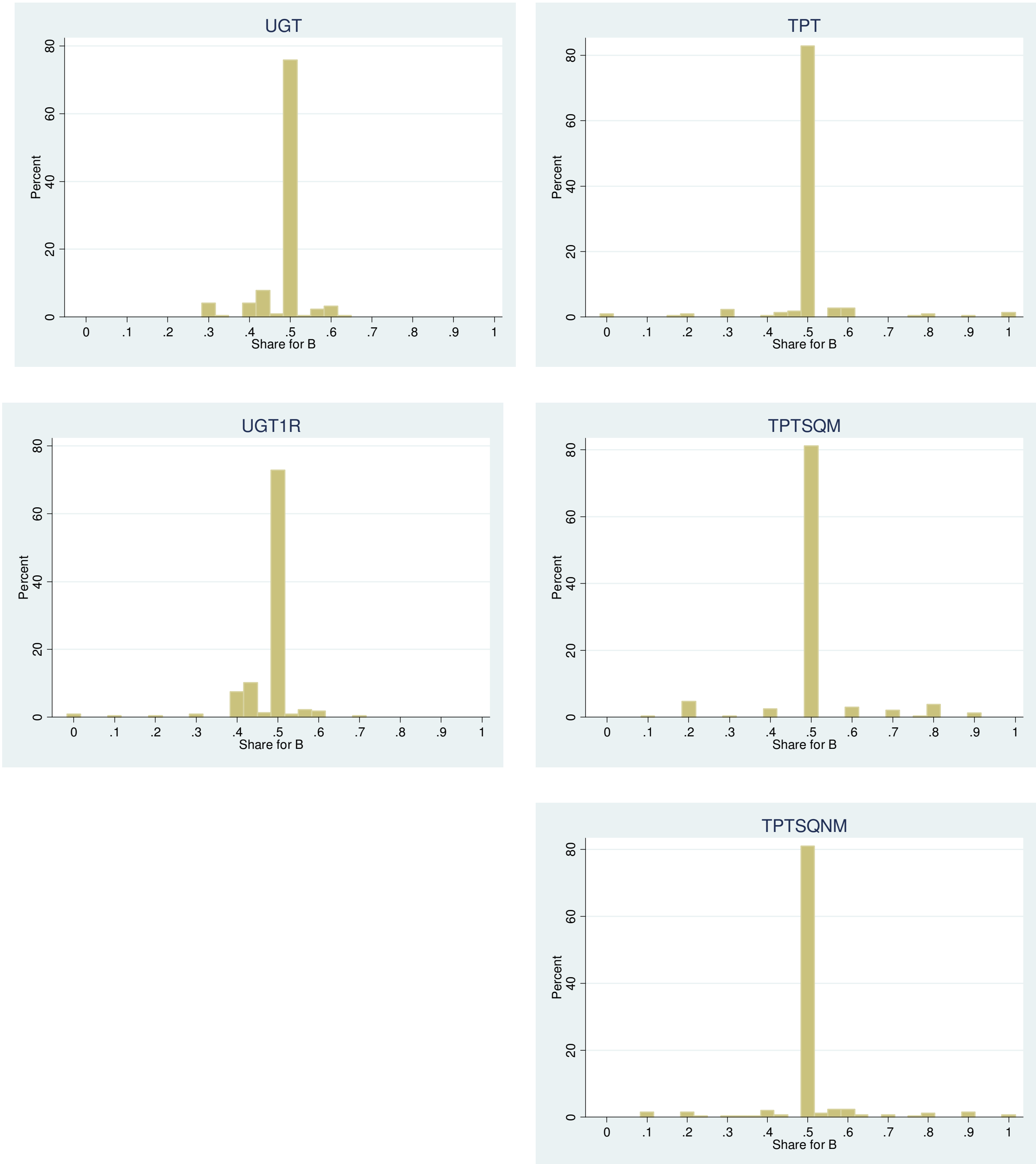


Figure 4: Frequencies (in percent) of shares that third parties and proposers distributed to subjects B / responders when both subjects gave the right answer (average frequencies for all nine matrices)

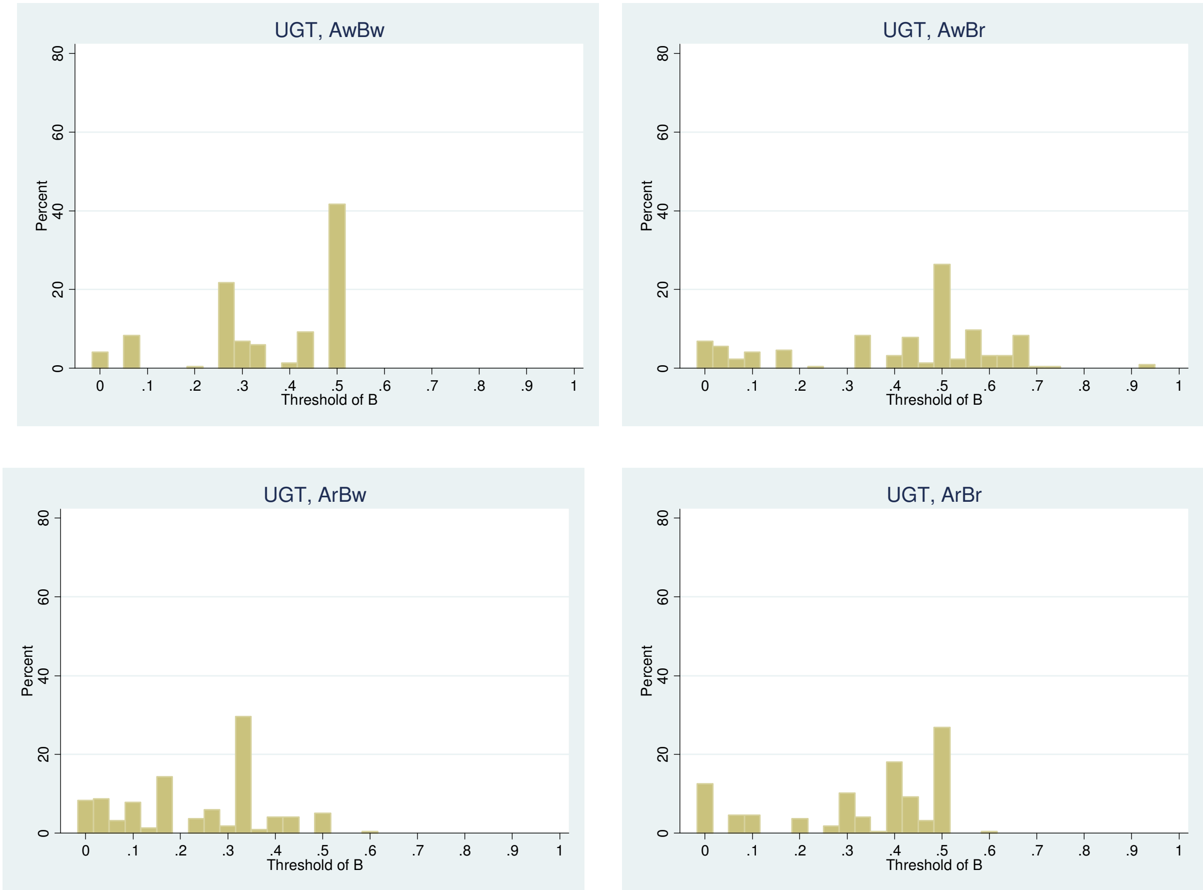


Figure 5: Frequencies (in percent) of responders' minimum claims for the cases $A_w B_w$, $A_w B_r$, $A_r B_w$, and $A_r B_r$ of the UGT (average frequencies for all nine matrices)

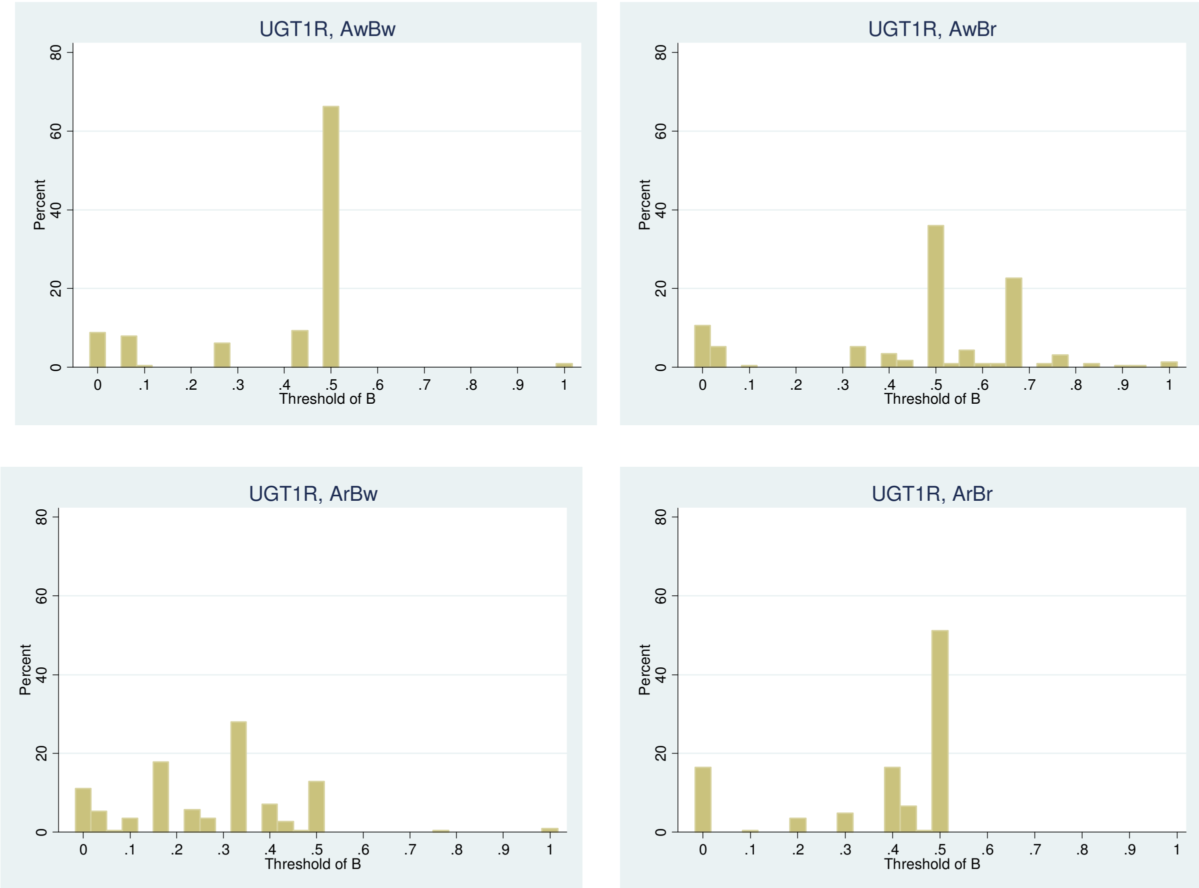


Figure 6: Frequencies (in percent) of responders' minimum claims for the cases $A_w B_w$, $A_w B_r$, $A_r B_w$, and $A_r B_r$ of the UGT1R (average frequencies for all nine matrices)

Literature

Alchian, Armen / Demsetz, Harold (1972): Production, information costs, and economic organization, in: American Economic Review 62 (5): 777-795.

Babcock, Linda / Loewenstein, George (1997): Explaining bargaining impasse: The role of self-serving bias, in: Journal of Economic Perspectives 11 (1): 109-126.

Babcock, Linda / Loewenstein, George / Issacharoff, Samuel / Camerer, Colin (1995): Biased judgments of fairness in bargaining, in: American Economic Review 85 (5): 1337-1343.

Baker, George P. / Jensen, Michael C. / Murphy, Kevin J. (1988): Compensation and incentives: Practice vs. theory, in: The Journal of Finance 43 (3): 593-616.

Bandiera, Oriana / Barankay, Iwan / Rasul, Imran (2007): Incentives for managers and inequality among workers: Evidence from a firm-level experiment, in: The Quarterly Journal of Economics 122 (2): 729-773.

Bolton, Gary E. / Ockenfels, Axel (2000): A theory of equity, reciprocity, and competition, in: American Economic Review 90 (1): 166-193.

Brandstätter, Hermann / Königstein, Manfred (2001): Personality influences on ultimatum bargaining decisions, in: European Journal of Personality 15 (1): 53-70.

Camerer, Colin / Thaler, Richard H. (1995): Anomalies: Ultimatums, dictators and manners, in: Journal of Economic Perspectives 9 (2): 209-219.

Cappelen, Alexander W. / Hole, Astri D. / Sørensen, Erik Ö. / Tungodden, Bertil (2007): The pluralism of fairness ideals: An experimental approach, in: American Economic Review 97 (3): 818-827.

Cappelen, Alexander W. / Moene, Karl O. / Sørensen, Erik Ö. / Tungodden, Bertil (2011): Rich meets Poor - An international fairness experiment, in: Journal of the European Economic Association forthcoming.

Cappelen, Alexander W. / Sørensen, Erik Ö. / Tungodden, Bertil (2010): Responsibility for what? fairness and individual responsibility, in: European Economic Review 54 (3): 429-441.

Charness, Gary / Haruvy, Ernan (2000): Self-serving biases: Evidence from a simulated labour relationship, in: Journal of Managerial Psychology 15 (7): 655-667.

Croson, Rachel T. A. (1996): Information in ultimatum games: An experimental study, in: Journal of Economic Behavior & Organization 30 (2): 197-212.

Croson, Rachel T. A. / Gomes, Armando G. / Mcginn, Kathleen L. / Nöth, Markus (2004): Mergers and acquisitions: An experimental analysis of synergies, externalities and dynamics, in: Review of Finance 8 (4): 481-514.

Croson, Rachel T. A. / Konow, James (2009): Social preferences and moral biases, in: Journal of Econnomic Behavior & Organization 69 (3): 201-212.

Dahl, Gordon B. / Ransom, Michael R. (1999): Does where you stand depend on where you sit? Tithing donations and self-serving beliefs, in: American Economic Review 89 (4): 703-727.

Doucouliagos, Hristos / Haman, Janto / Askary, Saeed (2007): Directors' remuneration and performance in Australian banking, in: Corporate Governance: An International Review 15 (6): 1363-1383.

Fehr, Ernst / Schmidt, Klaus M. (1999): A theory of fairness, competition, and cooperation, in: The Quarterly Journal of Economics 114 (3): 817-868.

Fischbacher, Urs (2007): Z-Tree: Zurich toolbox for ready-made experiments, in: Experimental Economics 10 (2): 171-178.

Gächter, Simon / Riedl, Arno (2005): Moral property rights in bargaining with infeasible claims, in: Management Science 51 (2): 249-263.

Gantner, Anita / Güth, Werner / Königstein, Manfred (2001): Equitable choices in bargaining games with joint production, in: Journal of Economic Behavior and Organization 46 (2): 209-225.

Greiner, Ben (2004): An online recruitment system for economic experiments, in: Forschung und wissenschaftliches Rechnen, edited by *Kremer, K. / Macho, V.*, Göttingen, Gesellschaft für Wissenschaftliche Datenverarbeitung. GWDG Bericht 63: 79-93.

Güth, Werner / Schmittberger, Rolf / Schwarze, Bernd (1982): An experimental analysis of ultimatum bargaining, in: Journal of Economic Behavior & Organization 3 (4): 367-388.

Hoffman, Elisabeth / Mccabe, Kevin / Shachat, Keith / Smith, Vernon A. (1994): Preferences, property rights, and anonymity in bargaining games, in: Games and Economic Behavior 7 (3): 346-380.

Izan, H. Y. / Sidhu, Baljit / Taylor, Stephen L. (1998): Does CEO pay reflect performance? Some Australian evidence, in: Corporate Governance: An International Review 6 (1): 39-47.

Kagel, John H. / Kim, Chung / Moser, Donald (1996): Fairness in ultimatum games with asymmetric information and asymmetric payoffs, in: Games & Economic Behavior 13 (1): 100-110.

Karagözoğlu, Emin / Riedl, Arno (2011): Information, uncertainty, and subjective entitlements in bargaining, IZA working paper #5079, retrieved: http://hdl.handle.net/10419/44224.

Königstein, Manfred (2000): Equity, efficiency and evolutionary stability in bargaining games with joint production, Berlin, Heidelberg, Springer-Verlag.

Konow, James (2000): Fair shares: Accountability and cognitive dissonance in allocation decisions, in: American Economic Review 90 (4): 1072-1091.

Konow, James (2003): Which is the fairest one of all? A positive analysis of justice theories, in: Journal of Economic Literature 41 (4): 1188-1239.

Konow, James (2009): Is fairness in the eye of the beholder? An impartial spectator analysis of justice, in: Social Choice and Welfare 33 (1): 101-127.

Konow, James / Saijo, Tatsuyoshi / Akai, Kenju (2009): Morals and mores: Experimental evidence on equity and equality, Mimeo, Loyola Marymount University, retrieved: http://www.dklevine.com/archive/refs4122247000000002055.pdf.

Kubo, Katsuyuki (2005): Executive compensation policy and company performance in Japan, in: Corporate Governance: An International Review 13 (3): 429-436.

Lazear, Edward P. (2000): Performance pay and productivity, in: American Economic Review 90 (5): 1346-1361.

Mitzkewitz, Michael / Nagel, Rosemarie (1993): Experimental results on ultimatum games with incomplete information, in: International Journal of Game Theory 22 (2): 171-198.

Oosterbeek, Hessel / Sloof, Randolph / Van De Kuilen, Gijs (2004): Cultural differences in ultimatum game experiments: Evidence from a meta-analysis, in: Experimental Economics 7: 171-188.

Paarsch, Harry J. / Shearer, Bruce S. (1999): The response of worker effort to piece rates: Evidence from the British Columbia tree-planting industry, in: Journal of Human Resources 34 (4): 643-667.

Pillutla, Madan M. / Murnighan, J. Keith (1995): Being fair or appearing fair: Strategic behavior in ultimatum bargaining, in: Academy of Management Journal 38 (5): 1408-1426.

Prendergast, Canice (1999): The provision of incentives in firms, in: Journal of Economic Literature 37 (1): 7-63.

Rajan, Madhav V. / Reichelstein, Stefan (2006): Subjective performance indicators and discretionary bonus pools, in: Journal of Accounting Research 44 (3): 585-618.

Rapoport, Amnon / Sundali, James A. (1996): Ultimatums in two-person bargaining with one-sided uncertainty: Offer games, in: International Journal of Game Theory 25 (40): 475-494.

Roth, Alvin E. (1995): Bargaining experiments, in: Handbook of Experimental Economics, edited by *Kagel, J. / Roth, A. E.*, Princeton, Princeton University Press: 253-348.

Selten, Reinhard (1967): Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments, in: Beiträge zur experimentellen Wirtschaftsforschung, edited by *Sauermann, H.*, Tübingen Mohr: 136-168.

Shapley, L. S. (1953): A value for N-person games, in: Contributions to the Theory of Games, edited by *Kuhn, H. / Tucker, A. W.*, Princeton, NJ, Princeton University Press. 2: 307-317.

Shiwakoti, Radha K. / Ashton, John K. / Keasey, Kevin (2004): Conversion, performance and executive compensation in UK building societies, in: Corporate Governance: An International Review 12 (3): 361-370.

Straub, Paul G. / Murnighan, J. Keith (1995): An experimental investigation of ultimatum games: Information, fairness, expectations, and lowest acceptable offers, in: Journal of Economic Behavior & Organization 27 (3): 345-364.

Van Dijk, Eric / De Cremer, David / Handgraaf, Michel J. J. (2004): Social value orientations and the strategic use of fairness in ultimatum bargaining, in: Journal of Experimental Social Psychology 40 (6): 697-707.

Chapter 2

How to Improve Patient Care? An Analysis of Capitation, Fee-for-Service, and Mixed Incentive Schemes for Physicians

Reference

Brosig-Koch, J., Hennig-Schmidt, H., Kairies, N., and D. Wiesen (2013a): "How to improve patient care? An Analysis of Capitation, Fee-for-Service, and Mixed Incentive Schemes for Physicians." *Ruhr Economic Papers* No. 412.

RUHR ECONOMIC PAPERS

Jeannette Brosig-Koch
Heike Hennig-Schmidt
Nadja Kairies
Daniel Wiesen

How to Improve Patient Care?

An Analysis of Capitation, Fee-for-Service, and Mixed Payment Schemes for Physicians

UNIVERSITÄT DUISBURG ESSEN

#412

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #412

Responsible Editor: Volker Clausen

ISSN 1864-4872 (online) – ISBN 978-3-86788-467-9

Ruhr Economic Papers #412

Jeannette Brosig-Koch, Heike Hennig-Schmidt,
Nadja Kairies, and Daniel Wiesen

How to Improve Patient Care?

An Analysis of Capitation, Fee-for-Service, and Mixed Payment Schemes for Physicians

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

http://dx.doi.org/10.4419/86788467
ISSN 1864-4872 (online)
ISBN 978-3-86788-467-9

Jeannette Brosig-Koch, Heike Hennig-Schmidt, Nadja Kairies, and Daniel Wiesen[1]

How to Improve Patient Care? – An Analysis of Capitation, Fee-for-Service, and Mixed Payment Schemes for Physicians

Abstract

In recent health care reforms, several countries have replaced pure payment schemes for physicians (fee-for-service, capitation) by so-called mixed payment schemes. Until now it is still an unresolved issue whether patients are really better off after these reforms. In this study we compare the effects resulting from pure and mixed incentives for physicians under controlled laboratory conditions. Subjects in the role of physicians choose the quantity of medical services for different patient types. Real patients gain a monetary benefit from subjects' decisions. Our results reveal that overprovision observed in fee-for-service schemes and underprovision observed in capitation schemes can, in fact, be reduced by mixed incentives. Interestingly, even the presentation of pure incentives as mixed incentives already significantly affects physicians' behavior. Moreover, the mixed payment schemes generally provide a higher benefit-remuneration ratio than the respective pure payment schemes. Our findings provide some valuable insights for designing health care reforms.

JEL Classification: C91, I11

Keywords: Physician incentive schemes; fee-for-service; capitation; mixed payment; laboratory experiment; presentation effect; benefit-remuneration analysis

April 2013

---

1 Jeannette Brosig-Koch, Nadja Kairies, Daniel Wiesen, University of Duisburg-Essen; Heike Hennig-Schmidt, University of Bonn. *– Financial support provided by the Deutsche Forschungsgemeinschaft (German Research Foundation) is gratefully acknowledged.* *– All correspondence to Jeannette Brosig-Koch, University of Duisburg-Essen, Faculty of Economics and Business Administration, Universitätsstr. 12, 45117 Essen, Germany, E-Mail: jeannette.brosig-koch@ibes.uni-due.de.*

1. Introduction

Traditionally, capitation (CAP) and fee-for-service (FFS) schemes were predominantly used to pay physicians in health care. With CAP physicians receive a fixed monetary amount for each patient assigned to them – irrespective of the quantity of medical care provided. FFS schemes pay a fixed amount of money for each medical service offered. Both payment schemes might incentivize physicians to deviate from the optimal treatment of their patients, however. While, with CAP, physicians are incentivized to provide less care than being optimal for their patients, FFS schemes embed an incentive to deliver more than the optimal level of medical service (see, e.g., Ellis and McGuire, 1986, or McGuire, 2000). This might result in detrimental effects for patients' health (see, e.g., McGuire, 2000). In order to counteract possible negative incentive effects for patients' health, recent reforms in the European and North American health care systems have replaced pure payment schemes by so-called mixed payment schemes or remuneration, i.e. schemes that combine CAP and FFS. As theoretically argued by Ellis and McGuire (1986), such mixed schemes can reduce the deficiencies of pure schemes.

Although, mixed payment schemes are often assumed as a 'cure' to the negative effect caused by pure incentives, comparisons of the consequences of both payment schemes, respectively, have received little attention in the empirical literature yet. Some studies suggest, at least, that patients are better off after this replacement using indirect measures to analyze the impact on patients' health benefit. For example, Krasnik et al. (1990) conducted a panel study in Denmark and found that general practitioners respond to a replacement of pure lump-sum payments by CAP supplemented by a FFS component by raising diagnostic and curative services and decreasing referrals to secondary care. Concerning referral rates, Iversen and Lurås (2000) obtain a similar result for Norway. They observe that referrals from primary to secondary care revealed by Norwegian general practitioners are larger under a CAP-scheme with a reduced FFS-component than under a scheme with a fixed payment (practice allowance component) complemented by a FFS-payment. The increase in referral decisions may, however, not only be attributable to CAP, but rather to the lower FFS-component. Dumont et al. (2008) analyze data on primary care services from the Canadian province Quebec before and after a variation from FFS to a mixed scheme with a base wage and a reduced FFS payment. Their results suggest that physicians respond to the mixed incentives by reducing the volume of services, but increasing the time spent per service and per nonclinical service. Also employing data from the Quebec payment reform, Échevin and Fortin (2011) analyze hospital patients' length of stay and risk of readmission. They find that the hospital length of stay of patients treated in departments under a mixed payment scheme increased substantially. Nevertheless, the risk of readmission with the same diagnosis does not appear to be overall affected by the reform.

Some field studies focusing on pure payment schemes find a rather weak or even no relationship between physicians' payment and their supply of medical services (see Gosden et al., 2001, or Sørensen and Grytten, 2003, for an overview). For example, Hutchinson et al. (1996) compare hospital utilization rates in Ontario (Canada) under FFS and CAP incentives and do not find any difference. Similarly, Hurley and Labelle (1995) conclude that the responses to pure payment incentives among Canadian physicians are rather mixed. After controlling for characteristics of patients and physicians, Grytten and Sørensen (2001) report that the impact of payment schemes on Norwegian physicians' behavior is rather small.

Since many of the field studies vary more than one component of the payment scheme simultaneously or might suffer from selection biases regarding patient characteristics, causal inferences on the direction and the strength of an effect are rather difficult. In addition, these studies are often based on self-reports which are not unlikely to differ from actual behavior (e.g., Camerer and Hogarth, 1999). In order to overcome some of the methodological deficiencies, Fuchs (2000) proposed to incorporate economic experiments as a complementary method to field studies in health economic research. Hennig-Schmidt et al. (2011) are among the first ones to follow this research agenda. They experimentally investigate the behavioral effects of FFS and CAP under controlled laboratory conditions. Their results demonstrate that these payment incentives significantly influence physician provision behavior. That is, they find support for the theoretically predicted underprovision with capitation and overprovision with fee-for-service, though patient benefits prove to be important as well.

Our study investigates the effects on physicians' provision behavior and consequences for patients' health that are associated with a replacement of pure payment incentives by mixed incentives. We base our study on controlled laboratory experiments, similar to Hennig-Schmidt et al. Our experimental design of pure payment incentives differs in three important aspects from the one of Hennig-Schmidt et al. (2011), however. First, our design allows directly comparing the two pure payment schemes and different mixtures of them with each other. Second, it allows identifying physician's responses to specific patient characteristics. Third, it allows assessing the effects of budget variations which often accompany a change of payment schemes. In contrast to field data, the experimental data allows measuring the impact of pure and mixed payment schemes directly on patients' health.

Our experiment particularly addresses the following questions: Do patients benefit from mixed incentives? That is, do mixed payment schemes mitigate over- and underprovision as predicted by theory (Ellis and McGuire, 1986)? Do the observed effects depend on specific patient characteristics? That is, how do the patient's illness and the severity of this illness affect the physician's behavior? Does it pay off for policy-makers in health care to implement mixed incentives? That is, does the 'patient benefit-physician remuneration' ratio improve with a mixed payment scheme?

The paper is organized as follows. Section 2 presents our experimental design and procedure. The results are provided in section 3. Section 4 summarizes our findings and concludes.

2. Experimental design and procedure

The aim of this study is to analyze the impact of different payment schemes on physicians' supply of medical services and on the patients' health benefits. Except for these schemes, no other experimental parameter is varied. The experiment, thus, allows for a controlled *ceteris paribus* analysis of the payment method.

In all experimental conditions subjects face the following decision situation: Each subject decides in the role of a physician and chooses a quantity of medical services for a given patient whose health benefit is influenced by that choice. More specifically, physician i decides on the quantity of medical services $q = \{0, 1, \dots, 10\}$ for nine patients $j \in [0,1,\dots,9]$. Patients differ in their illnesses $k \in [A, B, C]$ and in the severities $l \in [x, y, z]$ of these illnesses. The physician receives a certain payment which depends on the experimental condition (see below) and has to bear costs that depend on the quantity of medical services he or she chooses. Costs are assumed to be $c_{kl}(q) = 0.1 \cdot q^2$ in all conditions.[1] With each decision, the physician simultaneously determines her profit π_{kl}^{i} (payment R – cost c) and the patient's health benefit B_{kl}, both measured in monetary terms. The patient is assumed to be passive and fully insured, accepting each level of medical service provided by the physician.

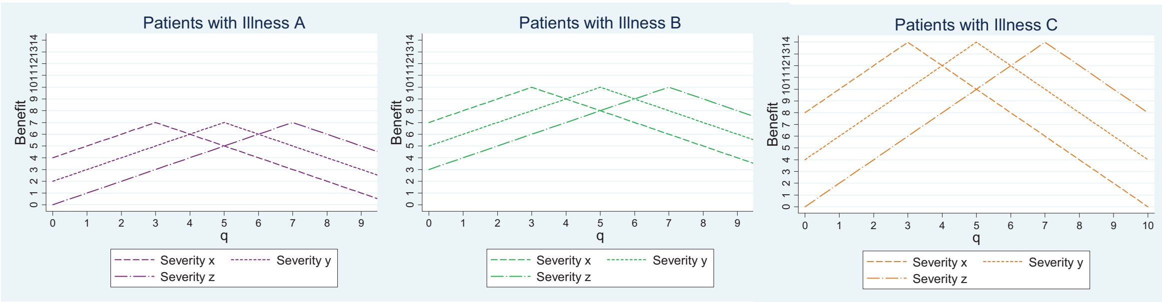
A common characteristic of the patient benefit functions is a global optimum on the quantity interval [0,10] (see Figure 1).[2] Illnesses A, B, C each imply a different level of health benefit. In particular, illnesses are modeled in a way that each illness is characterized by a different level of maximum health benefit $B_{Al}(q^*) = 7$, $B_{Bl}(q^*) = 10$, and $B_{Cl}(q^*) = 14$ and a certain slope of the benefit function (i.e., a certain change of benefit resulting from an additional unit of medical service). While the slope of the benefit function is the same for illnesses A and B, it is different for illness C. The optimal quantity q^* yielding patients the maximum health benefit $B_{kl}(q^*)$ from medical services depends only on the severity of an illness – moderate (x), intermediate (y) and severe (z). In our experiment, the patient's optimal quantities are $q^* = 3$ for severity *x*, $q^* = 5$ for severity *y*, and $q^* = 7$ for severity *z.* Taking q^* as the benchmark for the optimal medical treatment for the patient, we can identify overprovision and underprovision, respectively.[3] The optimal amount of medical services is specified for each patient and is known to the physician. Thus, there is no uncertainty about the impact of the chosen quantity of medical services on the patients' health benefit, and behavioral patterns like defensive medicine (see, e.g., Kessler and McCellan, 1996) can be neglected.

[1] In line with the theoretical literature, we assume convex cost functions (e.g., Ma, 1994, and Choné and Ma, 2010).

[2] A concave patient benefit function has been widely assumed in theoretical papers (see, e.g., Ellis and McGuire, 1986, Ma, 1994, Choné and Ma, 2010).

[3] From a medical point of view there might be several acceptable treatment variations among different physicians. This is not addressed in our simplified experimental setup. We, rather, assume that a specific amount of medical services renders the optimal health benefit to a patient.

Figure 1: Benefit functions for illnesses k and severities l



Patients are not actual subjects participating in the experiment. Real patients' health outside the lab is affected by decisions in the experiment, however. Subjects are informed accordingly, i.e., they know that the monetary equivalent to the patient benefit resulting from their decisions is transferred to a charity caring for real patients (the *Christoffel Blindenmission*, see also section 2.2.).

In order to study the change of payment schemes, we employ a within-subject design. That is, each subject participates in a session consisting of two parts. In part *I*, subjects decide under a pure payment scheme – either CAP or FFS. In part *II*, they decide under a mixed payment scheme (in the following labeled as MIX) – either with more weight on FFS or on CAP. Besides the within-subject comparison of pure and mixed incentive schemes (part *I* vs. part *II*), this design also allows for an across-subject comparison of the two pure incentive schemes in part *I* as well as of the different mixed incentives schemes in part *II*. In the following, the experimental conditions, i.e. payment schemes, are explained in more detail.

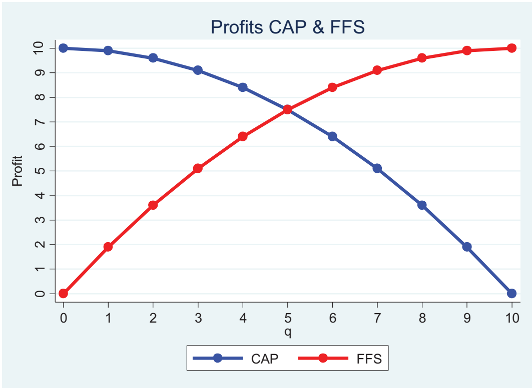
2.1. Experimental conditions

2.1.1. Pure payment schemes

Under CAP, each physician receives a lump-sum payment per patient of 10 regardless of illness k and severity l, i.e. $R = LS = 10$. Physician i's profit per patient is thus $\pi_{kl}^{i}(q) = 10 - c_{kl}(q)$. Under FFS, physicians are paid a fee of $p = 2$ per service they provide, i.e. $R(q) = 2q$ independent of illnesses k and severities l of an illness. Accordingly, physician i's profit per patient is $\pi_{kl}^{i}(q) = 2q - c_{kl}(q)$.

The maximum profit a physician can achieve is equal for both pure payment schemes, i.e. $\hat{\pi}_{kl}^{CAP} = \hat{\pi}_{kl}^{FFS} = 10$. Moreover, the marginal changes of profits are also the same in CAP and FFS. The only difference between the two schemes is the profit maximizing quantity of medical services $\hat{q}$, which is 0 for CAP and 10 for FFS. Quantity $\hat{q}$ does not depend on illness k and severities l of an illness. See Figure 2 for the profit functions in CAP and FFS (and Appendix B for the complete set of parameter values).

Figure 2: Profit functions in the pure payment schemes CAP and FFS



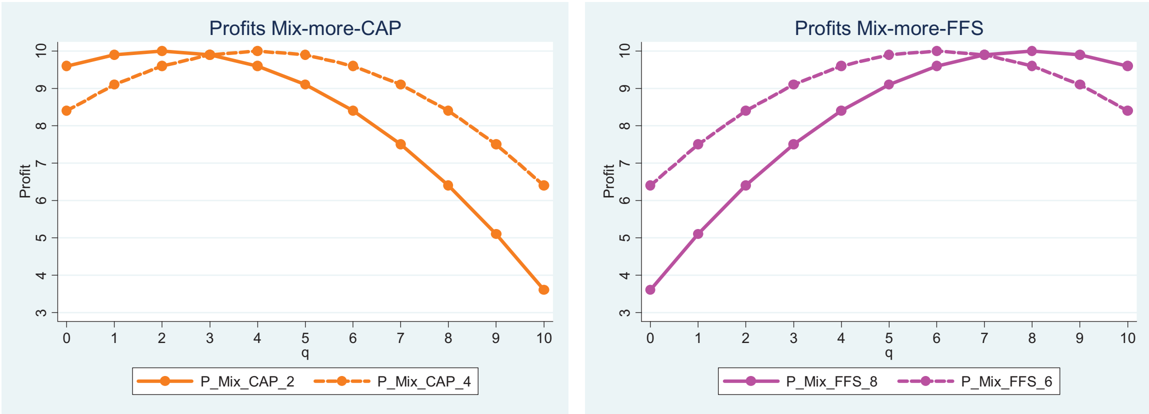
2.1.2. Mixed payment schemes

The mixed payment schemes comprise both a lump-sum and a fee-for-service component. More formally, $R_{kl} = \mu LS + (1-\mu)pq$. In the *MIX-more-CAP schemes,* we put more weight on the lump-sum component. To ensure that the maximum profit in MIX-more-CAP is the same as in the pure payment schemes (i.e., that physicians can earn as much under the pure as under the mixed schemes) the fee-for-service component p is adjusted. We implemented two schemes: MIX-more-CAP(2) and MIX-more-CAP(4). In MIX-more-CAP(2), the profit maximizing quantity $\hat{q}$ is 2, the weight of the lump-sum component μ is 0.96, and the fee per service p is 10. Accordingly, the payment for the physician is $R_{\text{MIX-more-CAP(2)}} = (0.96)10 + (0.04)10q$. In MIX-more-CAP(4), the values are $\hat{q} = 4$, $\mu = 0.84$, and $p = 5$, i.e., $R_{\text{MIX-more-CAP(4)}} = (0.84)10 + (0.16)5q$.

In the *MIX-more-FFS schemes*, the FFS component has a higher weight. To guarantee that the maximum profit is the same as in the MIX-more-CAP and pure payment schemes, we adjusted the lump-sum component LS. Again, we implemented two schemes: MIX-more-FFS(8) and MIX-more-FFS(6). In MIX-more-FFS(8), $\hat{q} = 8$, the weight of the FFS component $(1-\mu)$ is 0.80, and the lump-sum payment is 18. Accordingly, the payment for the physician is $R_{\text{MIX-more-FFS(8)}} = (0.20)18 + (0.80)2q$. In MIX-more-FFS(6), the values are $\hat{q}_{kl} = 6$, $(1-\mu) = 0.60$, and $LS = 16$, i.e. $R_{\text{MIX-more-FFS(6)}} = (0.40)10 + (0.60)2q$.

The profit functions of the MIX schemes are illustrated in Figure 3. By choosing $\hat{q} = 2$ and 4 in the MIX-more-CAP schemes and by choosing $\hat{q} = 6$ and 8 in the MIX-more-FFS schemes, we ensure that profit maxima are closer to the patient optima than in the pure payment schemes, but do not coincide with them. That is, we reduced the trade-off between profit-maximization and benefit maximization, though this trade-off does not vanish completely.

Figure 3: Profit functions in the mixed payment schemes

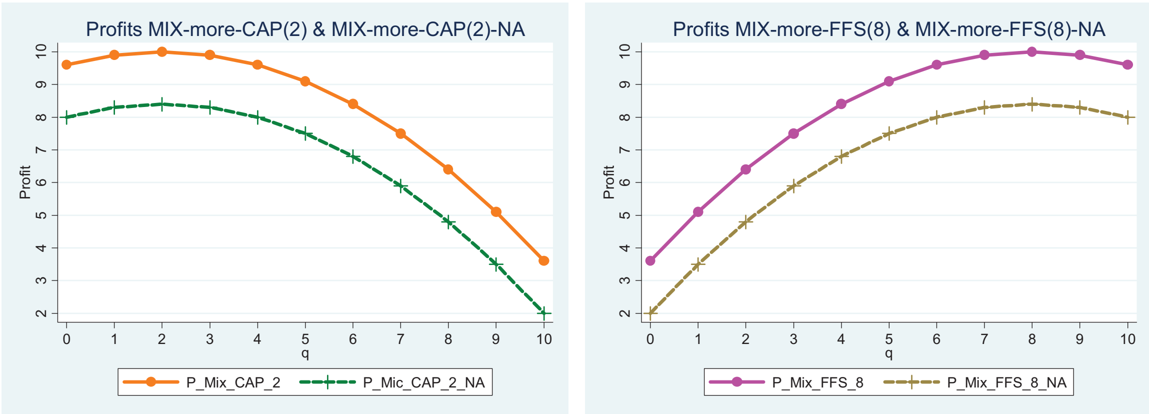


Keeping the maximum profit constant across pure and mixed payment schemes comes at the cost that the fee-for-service component and the lump-sum component have to be adjusted in the MIX schemes. Implementing the components from the pure schemes and using the same profit maximum quantities $\hat{q}_{kl}$ as in the MIX schemes would imply a lower profit maximum for the physician in the resulting 'non-adjusted' mixed payment schemes. That is, under these schemes it would not be possible for the physician to earn as much as under the pure payment schemes. This might significantly decrease physicians' incentives to consider the patient health benefit compared to the 'adjusted' mixed schemes.

From a perspective of a policy-maker not only the nature of physicians' payment schemes and the resulting incentives on physicians' behavior, but also the total cost for physicians' remuneration are relevant to judge upon the effectiveness of a reform of payment schemes. Accordingly, the total expenditures resulting from the different payment schemes have to be estimated. As long as we can assume that physicians are rational and purely selfish decision-makers, they will always choose the profit maximum and the total expenditure for the third-party payer is the same in the pure as in the 'adjusted' mixed schemes, while it is lower in the 'non-adjusted' mixed schemes. But if we assume, e.g., that physicians choose each possible quantity of medical service with equal probability, then the expected total expenditure in an 'adjusted' mixed scheme is higher than that in a pure scheme (which is equal to the one in the 'non-adjusted' mixed scheme).

In order to control for the incentive and expenditure effects that are associated with a lower profit maximum, we run two additional conditions with 'non-adjusted' mixed schemes, one with more weight on CAP and another with more weight on FFS. The two non-adjusted schemes are labeled MIX-more-CAP(2)-NA and MIX-more-FFS(8)-NA. Implementing these schemes should give more insights into the policy-makers' trade-off between introducing an effective mixed incentive scheme and keeping the total expenditure for physicians' payment constant. The scheme MIX-more-CAP(2)-NA is designed in a way that, first, both $R^{CAP} = LS = 10$ and $R^{FFS} = 2q$ from the pure payment schemes are included and, second, the physician's profit maximum is at $\hat{q}_{kl} = 2$ as in MIX-more-CAP(2). The weight on the lump-sum component μ is chosen such that these two criteria are met, i.e., $\mu = 0.80$. Also in scheme MIX-more-FFS(8)-NA, both $R^{CAP} = LS = 10$ and $R^{FFS} = 2q$ are included while the physician's profit maximum is at $\hat{q}_{kl} = 8$ as in MIX-more-FFS(8). The weight attached to the FFS component $(1 - \mu)$ is 0.80. The profit functions used in the two control conditions are illustrated in Figure 4. As this Figure reveals, the physician's maximum profit in the non-adjusted payment schemes is equal to 8.40 and, thus, lower than that in the pure and the non-adjusted mixed schemes (which is equal to 10).

Figure 4: Profit functions in the non-adjusted mixed payment schemes



2.1.3. Presentation and experience effects

In condition CAP-Presentation, we aim to test whether already the presentation of the payment scheme – either as a CAP or as a mixed scheme – influences physicians' behavior. That the presentation of monetary rewards alone can substantially affect behavior has been demonstrated in a number of studies (see, e.g., Pruitt 1967, 1970, Selten and Stoecker, 1986, Hannan et al., 2005, Gürerk and Selten, 2012, and Hossain and List, 2012).[4] For example, in a field experiment on an agricultural company, Englmaier et al. (2012) observe that a higher salience of incentives for quantity (i.e., the piece-rate) tends to increase the total amount of lettuce harvested. It is an open question whether such effects also translate to the presentation of physicians' payment schemes, however.

In order to isolate the effect of presentation, we design incentives, i.e., physicians' profit functions, in a way that they are exactly the same for both, part I and II. Only *how* the payment is presented to physicians (either as CAP or as mixed payment) differs between the two parts. This design allows us to identify the impact of the presentation on physicians' behavior at a within-subject level. To implement equal profits in a pure and a mixed scheme, we choose the cost function c_{kl}^{Pres} and the lump-sum payment per patient LS^{Pres} such that profits in the pure payment scheme CAP^{Pres} are exactly the same as those in the non-adjusted mixed payment scheme Mix-more-CAP(2)-NA. That is, in part I physician i

[4] See also the extensive literature on framing effects, i.e. effects that are caused by "*the decision-maker's conception of the acts, outcomes, and contingencies associated with a particular choice*" (Tversky and Kahneman, 1981, p. 453). While the seminal paper in this field was provided by Tversky and Kahneman (1981), more recent literature is surveyed by Kühberger (1998) and Levin et al. (1998).

receives $\pi_{kl}^{i}(q) = 15 - c_{kl}^{Pres}(q)$, with $c_{kl}^{Pres}(q) = 0.1q^2 - 0.4q + 7$. Costs c_{kl}^{Pres} are convex on the quantity interval $[0,10]$ with a minimum at 2 medical services.

In addition to the effect of presentation, we control for the effect of experience. That is, we control whether being confronted with a pure payment scheme in the part I affects decisions made in part II. Since the aim of our study is to test the replacement of pure payment schemes by mixed schemes, choosing a within-subject design (i.e., asking for their decisions in both payment schemes consecutively) is an appropriate approach.[5] Nevertheless, it might be interesting to isolate the behavioral effects which are due to the experience of subjects made in a pure payment scheme. Accordingly, in conditions MIX-more-CAP(4) and MIX-more-FFS (6), we implement the two mixed payment schemes without running a pure payment scheme beforehand. All experimental conditions and the number of participants are summarized in Table 1.

Table 1: Experimental conditions

Name	Condition	Part *I*	Part *II*	# Subjects
A-C2	CAP–MIX-more-CAP(2)	CAP	MIX-more-CAP(2)	22
A-C4	CAP–MIX-more-CAP(4)	CAP	MIX-more-CAP(4)	24
A-F8	FFS–MIX-more-FFS(8)	FFS	MIX-more-FFS(8)	24
A-F6	FFS–MIX-more-FFS(6)	FFS	MIX-more-FFS(6)	24
NA-C2	CAP–MIX-more-CAP(2)-NA	CAP	MIX-more-CAP(2)-NA	22
NA-F8	FFS–MIX-more-FFS(8)-NA	CAP	MIX-more-FFS(8)-NA	22
P-C2	CAP-Presentation	CAP^{Pres}	MIX-more-CAP(2)-NA	24
Ex-C4	MIX-more-CAP(4)	MIX-more-CAP(4)	–	23
Ex-F6	MIX-more-FFS (6)	MIX-more-FFS(6)	–	20

2.2. Experimental protocol

The computerized experiment was programmed with z-Tree (Fischbacher, 2007) and conducted at elfe, the Essen Laboratory for Experimental Economics at the University of Duisburg-Essen, Germany. Overall 205 students from the University of Duisburg-Essen participated in our experimental sessions. They were recruited by the online recruiting system ORSEE (Greiner, 2004).[6] Since we did not observe significant differences between decisions of medical students (who are supposed to become physicians in the future) and non-medical students in CAP and FFS, respectively (CAP/FFS: p=0.1880/0.1274, Fisher Pitman Permutation test for two independent samples (FPPI), two-sided), our subject pool is not restricted to students with a background in medicine.

[5] See, e.g., Kagel and Roth (2000) who use a similar approach to test the performance of centralized clearinghouse mechanisms.

[6] Students who registered in ORSEE to participate in laboratory experiments at the Essen Laboratory for Experimental Economics were invited via automatically generated e-mails and registered for a special session. We can thus say that subjects were randomly allocated to the experimental conditions. Moreover, subjects were not informed about the content of the experimental conditions unless they participated in a session.

The procedure was as follows: Upon arrival, subjects were randomly allocated to the cubicles. Then, they were given plenty of time to read the instructions for part *I* and to ask clarifying questions which were answered by the same experimenter in private. Subjects were informed that the experiment consisted of two parts, but received detailed instructions for part *II* only after having finished part *I* of the experiment. To check for subjects' understanding of the decision task, they had to answer a set of control questions. The experiment did not start unless all subjects had answered the control questions correctly. In each of the two parts of the experiment, subjects then subsequently decided on the quantity of medical service for each of the nine patients, i.e. for each possible combination of illnesses and severities. The order of patients was randomly determined and kept constant for all subjects in all conditions (see Table 2).

Table 2: Randomized order of illnesses and severities of illness

Patient *j*	1	2	3	4	5	6	7	8	9
Illness *k*	B	C	A	B	B	A	C	A	C
Severity *l*	*x*	*x*	*z*	*Y*	*z*	*y*	*z*	*x*	*y*

Before making their decision for a specific patient, subjects are informed about their payment, their cost and profit, as well as about the patient's benefit for each quantity from 0 to 10. All monetary amounts are given in Taler, our experimental currency unit, the exchange rate being 1 Taler = €0.08.[7] The procedure was exactly the same in part II of the experiment.

When all subjects had made their decisions, we randomly determined one decision in each part of the experiment to be relevant for a subject's actual payoff and the patient benefit. After the experiment, subjects were paid in private according to these two randomly determined decisions and were dismissed.

To verify that the money corresponding to the sum of patient benefits in a session was actually transferred, we applied a procedure similar to the one used in Hennig-Schmidt et al. (2011) and Eckel and Grossman (1996). To this end, one of the participants was randomly chosen to be the monitor. After the experiment, the monitor verified that an order on the aggregated benefit in the respective session was written to the financial department of the University of Duisburg-Essen to transfer the money to the *Christoffel Blindenmission*. The order was sealed in an envelope and the monitor and experimenter then walked together to the nearest mailbox and deposited the envelope. The monitor was paid an additional €5.

Sessions lasted for about 70 minutes. Subjects earned, on average, €15.74. The average benefit per patient was €12.94. In total, €2,652.70 were transferred to the *Christoffel Blindenmission*. The money supported surgical treatments of cataract patients in a hospital in Masvingo (Zimbabwe) staffed by ophthalmologists of the *Christoffel Blindenmission*.

[7] Instructions (including examples of the decision screen) and control questions are included in Appendix A.

Average costs for such an operation amounted to about €30. Thus, the money from our experiment allowed treating 88 patients.[8]

3. Results

3.1. Provision behavior under pure payment schemes

Before investigating the behavioral effects of introducing mixed incentive schemes, we compare physician's choices made in the two pure payment schemes of parts *I* of the MIX conditions. Aggregate data on decisions made in the pure payment schemes is included in Table 3.

Table 3: Descriptive statistics for behavior under pure payment schemes

Name	Condition	Part *I*	
		Mean	s.d.
A-C2	CAP–MIX-more-CAP(2)	3.04	2.08
A-C4	CAP–MIX-more-CAP(4)	3.44	1.90
A-F8	FFS–MIX-more-FFS(8)	7.57	2.34
A-F6	FFS–MIX-more-FFS(6)	7.19	1.94
NA-C2	CAP–MIX-more-CAP(2)-NA	2.78	1.96
NA-F8	FFS–MIX-more-FFS(8)-NA	6.89	2.10

In the experimental conditions A-C2, A-C4, and NA-C2, we used the same CAP payment scheme in part *I* of the experiment. Since physicians' choices under this payment scheme do not differ significantly between the three conditions either on the aggregate level (p>0.1398) or on the patient level, except for three of the 3x9 comparisons (A2-C2/A-C4: patient 3 p=0.0900, A-C2/NA-C2: patient 1 p=0.0480, A-C4/NA-C2: patient 7 p=0.0520, FPPI, two-sided), we pool data over these CAP conditions when analyzing the incentive effects of the two pure payment schemes. Similarly, in conditions A-F8, A-F6, and NA-F8, we used the same FFS payment scheme in part *I*. Since physicians' choices under FFS do not differ significantly between these conditions either on the aggregate level (p>0.2140) or at the patient level except for one of the 3x9 comparisons (A-F8/NA-F8: patient 9 p=0.0686, FPPI, two-sided), we also pool data over these FFS conditions in our analysis of pure payment schemes.

Comparing physicians' quantity choices between the CAP and the FFS payment schemes for all three illnesses k and severities l reveals that physicians do respond to financial incentives. We particularly observe that patients are underprovided in CAP and that they are overprovided in FFS (see Figures 5 and 6). That is, the quantity chosen by the physician is lower in CAP and higher in FFS than the quantity yielding the maximum health benefit for the patient. This effect is significant for all patients (p<0.0002, Wilcoxon-signed-rank-test, two-sided).

[8] Subjects were not informed about the money being assigned to a developing country to avoid motives like compassion.

Figure 5: Benefit functions, profit functions, and average quantities per severity in CAP

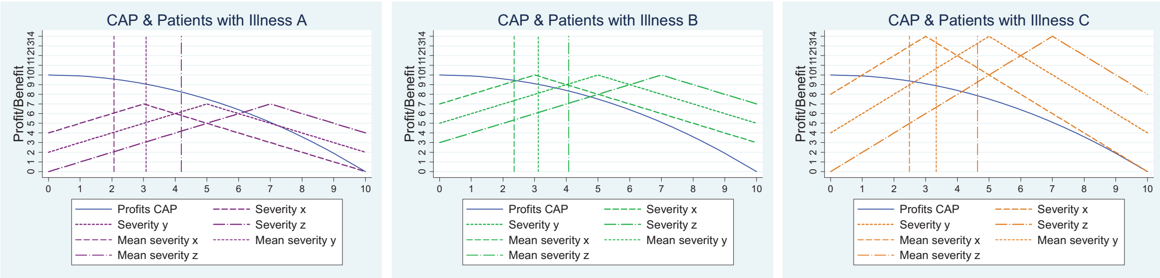
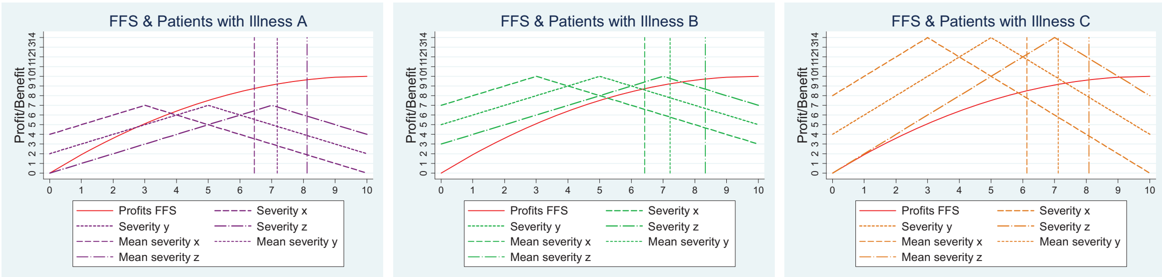
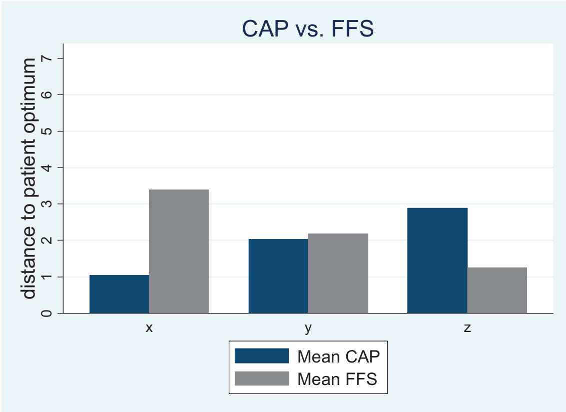


Figure 6: Benefit functions, profit functions, and average quantities per severity in FFS



Our results, thus, confirm the significant underprovision with capitation and the significant overprovision with fee-for-service observed for different parameters by Hennig-Schmidt et al. (2011). Our results, thus, also relate to earlier field studies by Gaynor and Gertler (1995) or by Gaynor and Pauly (1990). However, we extend the findings on pure payment schemes insofar as we systematically vary benefit functions representing certain characteristics of a patient. Accordingly, we are able to associate the degree of underprovision and overprovision, respectively, with these patient characteristics. In particular, we find that, under both pure payment schemes and in all conditions, physicians' decisions do significantly respond to the severity of an illness ($p<0.0054$), but hardly to the illness itself ($p>0.1049$, except for two of the 3x2x3 comparisons where $p<0.0135$, Fisher Pitman Permutation test for paired replicates (FPPP), two-sided). That is, neither the level of patient health benefit that could be maximally realized by the physician, nor the increase (and decrease, respectively) of health benefit that results from an additional unit of medical service (both are implied by an illness; see Section 2.1.1) systematically affects physicians' behavior. Only the quantity yielding the maximum health benefit (which is implied by the severity of an illness) clearly influences the choice of medical services. The more the optimum quantity for the patient deviates from the profit maximizing quantity for the physician, the more underprovision in CAP and overprovision in FFS, respectively, is observed. More specifically, calculating the distance between the quantity chosen by the physician and the quantity that is optimal for the patient we find that this distance significantly increases the more severe the illness is in CAP and the less severe the illness is in FFS ($p<0.0000$, FPPP, two-sided). Figure 7 illustrates these findings. In the following, we pool physicians' decisions over the three illnesses A, B, C.

Figure 7: Distance between patient optimum and chosen quantity per severity



Finally, due to the symmetry of our pure payment schemes we are able to test whether incentives to underprovide in CAP are equally strong as incentives to overprovide in FFS. From a psychological point of view, people might regard choosing a higher quantity of medical services for the patient (which is incentivized in FFS) less severe than omitting this quantity from the patient (which is incentivized in CAP) - though, in our experiment the loss of benefit for the patient is the same in both cases. Accordingly, the problem of overprovision in FFS might be more severe than the problem of underprovision in CAP. Comparing the distance between the quantity chosen by the physician and the quantity that is optimal for the patient between the two pure payment schemes does not support this supposition, however (CAP: 1.99 vs. FFS: 2.28; p=0.2754, FPPI, two-sided).[9]

Result 1 (Provision behavior under pure payment schemes)
We observe significant underprovision with CAP and significant overprovision with FFS. The deviations from the patient optimal quantity of medical services observed with the two payment schemes are equally severe. Under both schemes, the severity of an illness has a significant and systematic effect whereas the illness itself does not.

3.2. Comparison of pure and mixed payment schemes

Effects from changes of the payment method from a pure to a mixed scheme on physician provision behavior can be identified by our within-subject design. Aggregate data on decisions in the mixed payment schemes is included in Table 4. Note that, also in the mixed schemes, physicians' decisions hardly ever respond to the patient's illness (p>0.1049, except for A-F6: A vs. B p=0.0017 and B vs. C p=0.0134), but they do significantly so to its severity (p<0.0053, FPPP, two-sided). Accordingly, we pool physicians' decisions in the mixed payment schemes over the three illnesses A, B, C for all subsequent analyses.

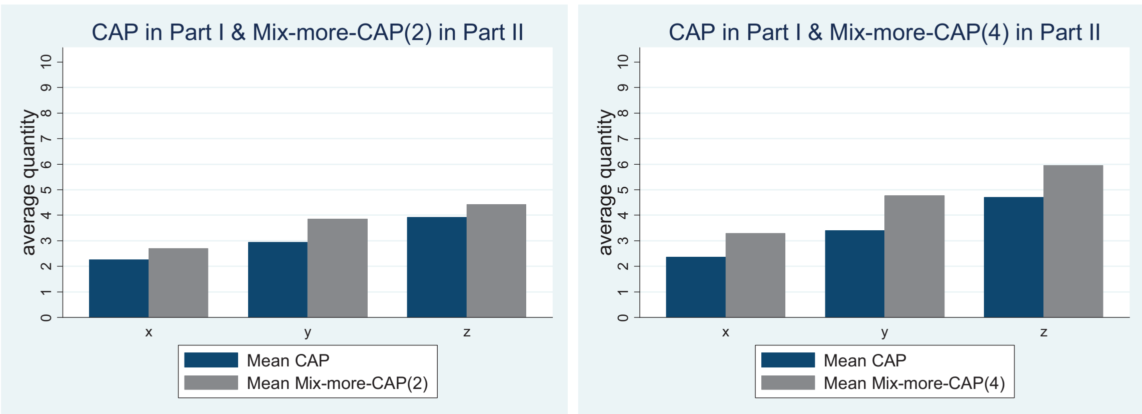
[9] Symmetry for the distance to the patient optimum holds also for comparisons of Mix-more-CAP(2) with Mix-more-FFS(8), and Mix-more-CAP(4) with Mix-more-FFS(6). Also between the symmetric mixed incentive schemes we find no significant difference regarding this distance (p>0.2283, FPPI, two-sided).

Table 4: Descriptive statistics for behavior under mixed payment schemes

Name	Condition	Part *II*	
		Mean	s.d.
A-C2	CAP–MIX-more-CAP(2)	3.66	1.74
A-C4	CAP–MIX-more-CAP(4)	4.65	1.32
A-F8	FFS–MIX-more-FFS(8)	6.92	1.73
A-F6	FFS–MIX-more-FFS(6)	5.62	1.21
NA-C2	CAP–MIX-more-CAP(2)-NA	3.46	1.43
NA-F8	FFS–MIX-more-FFS(8)-NA	6.50	1.53

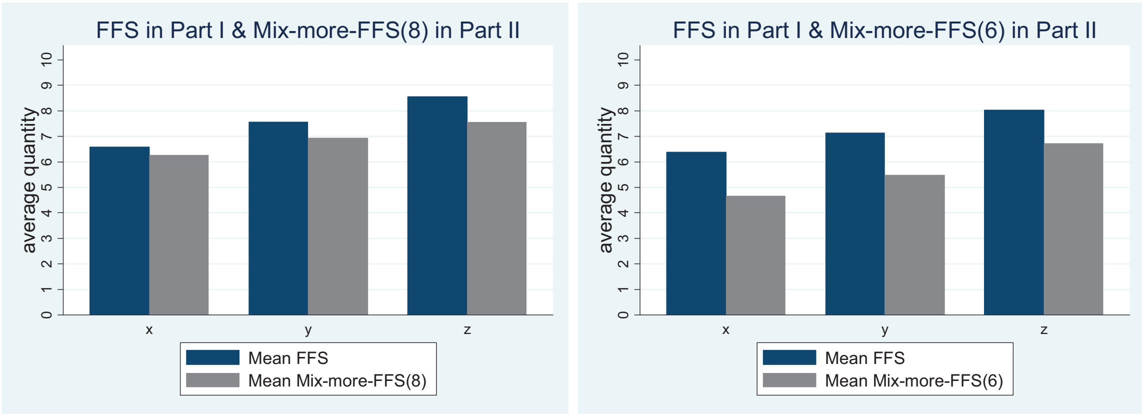
Comparing behavior of physicians in the pure CAP payment scheme (in part *I)* with their behavior in the respective (adjusted) MIX-more-CAP scheme (in part *II)* yields a significantly higher provision of services in the latter (A-C2: over all severities p=0.0017; differentiated per severity p≤0.0000; A-C4: average over all severities p=0.000; differentiated per severity: p≤0.0416, FPPP, two-sided). That is, introducing a mixed payment scheme that yields the same profit maximum as in CAP significantly reduces the underprovision observed in CAP. This effect increases with decreasing lump-sum component in the mixed payment scheme (0.96 in MIX-more-CAP(2) vs. 0.84 in MIX-more-CAP(4); p=0.0002, FPPI, two-sided). Though, in both mixed schemes and for all severities there is still a significant deviation from the quantity of medical services that is optimal for the patient (p<0.0040, Wilcoxon signed rank test, two-sided). The results are illustrated in Figure 8.

Figure 8: Average quantity choices in CAP and MIX-more-CAP payment schemes



Comparing behavior of physicians under FFS in part *I* with their behavior in the respective (adjusted) MIX-more-FFS scheme in part *II* yields a significantly lower quantity of medical service in the latter (A-F6: over all severities p=0.0082; differentiated per severity p≤0.0000; A-F8: average over all severities p=0.000; differentiated per severity: *x*: p=0.2436, *y*: p=0.0299, *z*: p=0.0003, FPPP, two-sided). That is, introducing a mixed incentive scheme that yields the same profit maximum for the physician as in FFS significantly reduces the overprovision observed in FFS. This effect is stronger the less weight is given to the FFS component in the mixed payment scheme (0.80 in MIX-more-FFS(8) and 0.60 in MIX-more-FFS(6); p=0.0004, FPPI, two-sided). Though, in both mixed schemes and for all severities there is still a significant deviation from the quantity of medical services that is optimal for the patient (p<0.0038, FPPP, two-sided). The results are illustrated in Figure 9.

Figure 9: Average quantity choices in FFS and MIX-more-FFS payment schemes

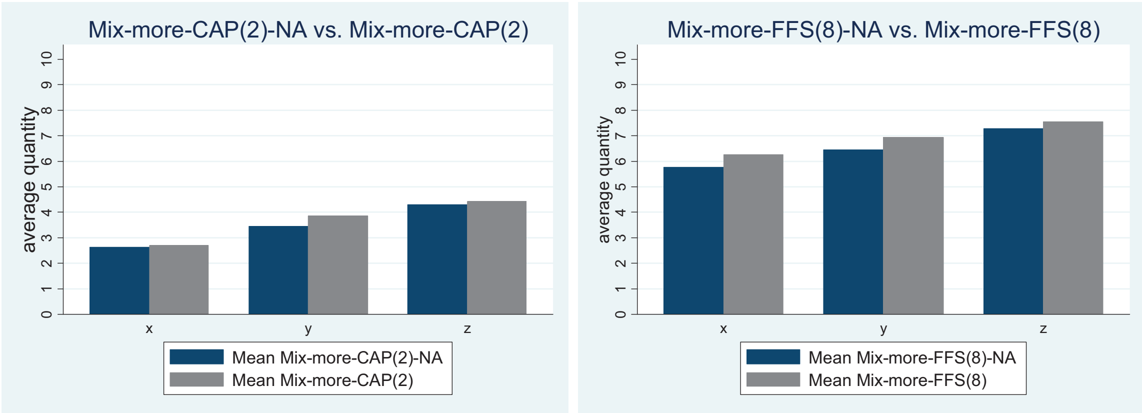


Result 2 (Impact of the introduction of mixed payment schemes)
Introducing mixed incentive schemes significantly reduces the underprovision observed with CAP and the overprovision observed with FFS, respectively.

3.3. Analysis of non-adjusted mixed schemes: Incentive and expenditure effects

In order to ensure that a physician can earn as much under the pure as under the mixed payment schemes, we adjusted the fee-for-service component and the lump-sum component in the mixed schemes tested in conditions A-C2, A-C4, A-F6, and A-F8, respectively. Without this adjustment, the physician's profit maximum (and, correspondingly, the physicians' monetary incentives) would be lower in the mixed schemes than in the pure payment schemes. In order to control for the incentive and expenditure effects that are associated with a lower profit maximum, we compare behavior in the adjusted and non-adjusted mixed schemes (i.e., MIX-more-CAP(2) vs. MIX-more-CAP(2)-NA and MIX-more-FFS(8) vs. MIX-more-FFS(8)-NA). Interestingly, we find neither a significant effect for the MIX-more-CAP schemes (p=0.5804, all FPPI, two-sided) nor for the MIX-more-FFS schemes (p=0.2886, all FPPI, two-sided). That is, reducing the maximal payoff physicians can achieve in the mixed payment schemes does not affect the quantity of medical services provided. Accordingly, also when introducing the two non-adjusted mixed payment schemes physicians tend to reduce their deviation from the patient's optimal quantity compared to the pure payment schemes (CAP vs. MIX-more-CAP(2)-NA: p=0.0001; differentiated per severity p≤0.0004, FFS vs. FFS–MIX-more-FFS(8)-NA: p=0.03206; differentiated per severity x: p=0.3256, y: p=0.1893, z: p=0.0268; FPPP, two-sided). Figure 10 illustrates these findings.

Figure 10: Average quantity choices in the non-adjusted and the adjusted mixed schemes



Looking at the third-party's total expenditure for physicians' payment (i.e., their remuneration) in the adjusted and non-adjusted mixed payment schemes, we find that this is significantly higher in the former than in the latter (MIX-more-CAP(2)-NA vs. MIX-more-CAP(2): p=0.0000 and MIX-more-FFS(8)-NA vs. MIX-more-FFS(8) p=0.0007, FPPI, two-sided; see Table 5 below). But how do the expenditures that result in the adjusted and non-adjusted mixed schemes relate to the ones that result in the pure payment schemes? Comparing total expenditures for physicians payment between the non-adjusted mixed and the pure schemes we find significantly lower expenditures in the former than in the latter (NA-C2: p = 0.0000, NA-F8: p = 0.0018, FPPP, two-sided). The adjusted mixed schemes with a larger CAP component yield significantly higher expenditures than CAP (A-C2: p=0.0000; A-C8: p=0.0000), while the adjusted mixed schemes with a larger FFS component yield significantly lower expenditures than FFS, if at all (A-F8: p=0.3344, A-F6: p=0.02573).

Result 3 (Provision behavior under non-adjusted mixed payment schemes)
A lower maximum profit for physicians does not significantly affect physicians' behavior, but yields a lower expenditure for physicians' payment in the non-adjusted mixed schemes (which is also lower than the expenditure in the respective pure payment schemes).

Although total expenditures in the adjusted Mix-more-CAP schemes are higher than those in the pure schemes, from a perspective of a policy-maker both, the remuneration for physicians and the benefit for patients, are important to judge the effectiveness of a reform of payment schemes. Result 3 implies that keeping the nature of incentives constant suffices to achieve a physician behavior similar to a high-powered incentive scheme resulting in higher expenditures for the third-party payer. The next section provides an in-depth benefit-remuneration analysis of the pure and the mixed payment schemes.

3.4. Benefit-remuneration analysis of pure and mixed payment schemes

When reforming physician payment schemes, third-party payers often rely on cost benefit analyses of the impact of reforms on patient welfare and health care costs (see, e.g., Garber, 2000). In field studies, prominent measures for patients' health benefit have been, among others, the time spent with a patient, the number of patient visits, or the referrals to specialists (see, e.g., Dumont et al., 2008). Obviously, implications for the patients' health benefit are rather indirect, rendering a full-fledged analysis of patients' welfare difficult in the field. The behavioral data from our experiment are suitable for the analysis of third payers' costs as our design allows to control and to directly measure variations in physicians' remuneration and corresponding changes in patients' health benefit. Although, admittedly, the experimental setting is stylized, it gives important insights into the relative performance of pure and mixed payment schemes resulting from physicians' treatment behavior.

Our results in the last sections reveal that the health benefit of an average patient is higher in a mixed payment scheme than in the corresponding pure payment scheme (see Table 5). For the cost of physicians' payment, the picture resulting from a comparison of pure and mixed schemes is less clear-cut (see section 3.3.). In particular, physicians' remuneration is lower in the adjusted mixed schemes with a larger FFS component and in both non-adjusted mixed schemes than in the corresponding pure schemes, while in the adjusted mixed schemes with a larger CAP component the remuneration is higher than in CAP.

Table 5: Descriptive statistics on patients' health benefit and physicians' remuneration by payment scheme

	Part *I* (*pure payment schemes*)		Part *II* (*mixed payment schemes*)	
Condition	Avg. Patient Benefit	Avg. Remuneration	Avg. Patient Benefit	Avg. Remuneration
CAP–MIX-more-CAP(2)	7.42	10.00	8.27	11.06
CAP–MIX-more-CAP(4)	8.31	10.00	9.64	12.12
FFS–MIX-more-FFS(8)	6.84	15.14	7.68	14.67
FFS–MIX-more-FFS(6)	7.47	14.38	9.26	13.15
CAP–MIX-more-CAP(2)-NA	7.34	10.00	8.27	9.38
FFS–MIX-more-FFS(8)-NA	7.72	13.78	8.35	12.41

In the following, we analyze the average ratios of patient benefit and physician's remuneration for both the pure and the mixed payment schemes (see Table 6). Replacing a pure scheme by a mixed scheme usually improves the benefit-remuneration ratio. This effect is significant for the replacement of a fee-for-service payment scheme by an adjusted mixed scheme with a relatively high weight on the lump-sum component (0.40) and by a non-adjusted mixed scheme (FFS vs. MIX-more-FFS(6): p=0.0280, FFS vs. MIX-more-FFS(8)-NA: p=0.0891) and for replacing a capitation payment scheme by a non-adjusted mixed scheme (CAP vs. MIX-more-CAP(2)-NA: p=0.0000).

Comparing the benefit-remuneration ratio between the two pure incentive schemes, we find a significantly higher ratio in CAP than in FFS (p=0.0027, FPPI, two-sided). Similarly, also the mixed schemes with a higher weight on the lump-sum component imply a significantly higher ratio than the schemes with a higher weight on the FFS component (p=0.0027, FPPI, two-sided). That is, compared over all payment schemes, we observe the highest benefit-remuneration ratio in the MIX-more-CAP(2)-NA scheme.

Table 6: Analysis of patients' health benefit and physicians' remuneration ratio

Condition	Part I *(pure payment schemes)* Avg. (Benefit/ Remuneration)	Part II *(mixed payment schemes)* Avg. (Benefit/ Remuneration)
CAP–MIX-more-CAP(2)	0.74	0.75
CAP–MIX-more-CAP(4)	0.83	0.80
FFS–MIX-more-FFS(8)	0.60	0.57
FFS–MIX-more-FFS(6)	0.61	0.71
CAP–MIX-more-CAP(2)-NA	0.73	0.88
FFS–MIX-more-FFS(8)-NA	0.68	0.73

Result 4 (Ratio of health benefit and remuneration)
Almost all mixed payment schemes yield a higher benefit-remuneration ratio than the respective pure payment schemes. Payment schemes comprising a CAP-component attain the highest values.

Taken at its face value, our results render for a third-party payer who is interested in the ratio of benefits and remuneration the non-adjusted Mix-more-CAP scheme most attractive. Naturally, the lowest remuneration for the average physician observed in this scheme contributes to the favorable benefit-remuneration ratio. A third-party payer giving more weight to the patient health benefit might opt for an adjusted Mix-more-CAP scheme with a higher weight on the lump-sum component instead.

3.5. Presentation and experience effects

Finally, we test whether the presentation of physicians' profit as the result of a pure payment scheme (part *I*) or as the result of a mixed scheme (part *II*) already influences behavior. Our results reveal, in fact, significant differences regarding this presentation: physicians choose a significantly higher quantity of medical service if the profit results from a mixed payment scheme than if it results from a pure payment scheme (p=0.01741, FPPP, two-sided). Differentiating according to severities, we find this effect to be due to a highly significant difference for severity *y* only (*x*: p=0.3282; *y*: p=0.0012; *z*: p=0.2452, all FPPP, two-sided).

In order to find out to what extend the reported benefit-improving effects of the mixed incentive schemes are due to the experience of subjects made with the pure schemes in part *I* of the experiment, we repeated two of the mixed payment schemes without a first part (MIX-more-CAP(4) and MIX-more-FFS (6)). Comparing the two mixed schemes with and without part *I* of the experiment, we find no significant differences (p=0.9638 and p=0.1243, all FPPI, two-sided). That is, subjects' experience does not alter their behavior in part *II* of the experiment, at least not to a significant degree.

Result 5 (Presentation and experience)
While the presentation of a profit function as being the result of a mixed payment scheme significantly improves patient benefits, the experience of subjects with a pure payment scheme does not significantly affect behavior.

4. Discussion and Concluding Remarks

This study provided a laboratory test of the effects that are associated with a health care provider payment reform – i.e., the replacement of pure payment incentives by mixed incentives. Regarding the pure incentive schemes, our data support the significant underprovision with capitation payment and significant overprovision with fee-for-service payment as predicted by theory (Ellis and McGuire, 1986) and suggested by results from previous studies (see, e.g., Hennig-Schmidt et al., 2011). In addition to previous research, our systematic variation of patient characteristics – illness and severity of illness – reveals that neither the level of patient health benefit that could be maximally realized by the physician, nor the increase (and decrease, respectively) of health benefit that results from an additional unit of medical service systematically affects physicians' decisions. Only the quantity yielding the maximum health benefit clearly influences the choice of medical services. The more the optimum quantity for the patient deviates from the profit maximizing quantity for the physician, the more underprovision in a capitation payment scheme or overprovision in a fee-for-service payment scheme, respectively, is observed.

Combining capitation with fee-for-service incentives, we find that the significant underprovision and overprovision of medical services observed with CAP and FFS, respectively, can be mitigated significantly. Even if the supplemented pure scheme receives relatively little weight, we observe a significant increase of the quantity of medical service provided by physicians. Accordingly, our experimental data - providing a *direct* measure for the patients' health benefit - reveal that patients experience a higher health benefit in mixed payment schemes compared to pure payment schemes. Interestingly, our finding relates to evidence from field studies arguing that patient are better off after the change from a pure to a mixed scheme (e.g., Krasnik et al., 1990, or Iversen and Luras, 2000). Moreover, presenting physician's profit as the result of a mixed incentive scheme instead of a capitation scheme already significantly increases physicians' care for the patient. That is, mixing pure payment schemes with each other positively affects provision behavior beyond the mere monetary incentives.

Our analysis of total expenditures for physicians' payment demonstrates that it is possible to design mixed schemes that *decrease* the total payment for physicians, but still reduce the physicians' deviation from the patient optimal quantity of medical care. These results are appealing from a welfare economics perspective. Calculating the ratio of patient benefits and physicians' remuneration we find that our non-adjusted mixed scheme with a relatively high weight on the capitation component yields the highest benefit-remuneration ratio.

As such, our experimental study provides valuable implications for healthcare reforms that include the introduction of mixed payment schemes for physicians. A policy maker or a third-party payer focusing on a well-balanced ratio between expenditures for physicians' remuneration and patients' health benefit would favor a mixed capitation scheme that does not adjust physicians' maximum payment to the level of the previous pure capitation payment scheme. A third-party payer interested in the patient health benefit would opt for an adjusted capitation scheme with a high weight on the lump-sum component instead.

References

Camerer, C. F., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*(1), pp. 7-42.

Choné, P., & Ma, C. A. (2010). Optimal Health Care Contracts under Physician Agency. *Working Paper, Boston University*.

Dumont, E., Fortin, B., Jacquement, N., & Shearer, B. (2008). Physicians' Multitasking and Incentives: Empirical Evidence from a Natural Experiment. *Journal of Health Economics, 27*, pp. 1436-1450.

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

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

Englmaier, F., Roider, A., & Sunde, U. (2012). The role of salience in performance schemes: Evidence from a field experiment. *Working Paper*.

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

Fuchs, V. R. (2000). The future of health economics. *Journal of Health Economics, 19*(2), pp. 141-157.

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

Greiner, B. (2004). An Online Recruitment System for Economic Experiments. (K. Kremer, & V. Macho, Eds.) *Forschung und wissenschaftliches Rechnen : Beiträge zum Heinz-Billing-Preis 2003*, pp. 79-93.

Grytten, J., & Sørensen, R. (2001). Type of Contract and Supplier-Induced Demand for Primary Physicians in Norway. *Journal of Health Economics, 20*, pp. 379-393.

Gürerk, Ö., & Selten, R. (2012). The effect of payoff tables on experimental oligopoly behavior. *Experimental Economics*, pp. 1-11.

Hannan, R., Hoffman, V., & Moser, D. (2005). Bonus versus penalty: does contract frame affect employee effort? *Experimental Business Research*, pp. 151-169.

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

Hossain, T., & List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science, fothcoming*.

Hurley, J., & Labelle, R. (1995). Relative fees and the utilization of physicians' services in Canada. *Health Economics, 4*(6), pp. 419-438.

Hutchinson, B., Birch, S., J. H., Lomas, J., & Stratford-Devai, F. (1996). Do Physician Payment Mechanisms Affect Hospital Utilization? A Study of Health Service Organizations in Ontario. *Canadian Medical Association Journal, 154*, pp. 653-661.

Iversen, T., & Lurås, H. (2000). The effect of capitation on GPs' referral decisions. *Health Economics, 9*(3), pp. 199-210.

Kessler, D., & McClellan, M. (1996). Do Doctors Practice Defensive Medicine? *Quarterly Journal of Economics, 111*, pp. 353–390.

Krasnik, A. P. (1990). Changing Remuneration Systems: Effects on Activity in General Practice. *British Medical Journal, 300*, pp. 1698-1701.

Kühberger, A. (1998). The influence of framing on risky decisions: a meta-analysis. *Organizational Behavior and Human Decision Processes, 75*, pp. 23–55.

Levin, I. P., Schneider, S., & Gaeth, G. J. (1998). All frames are not created equal: a typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes, 76*, pp. 149-188.

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

McGuire, T. G. (2000). Physician Agency. In *Handbook of Health Economics, Vol. 1 A.* Amsterdam, North-Holland: A. C. Cuyler and J. P. Newhouse.

Pruitt, D. G. (1967). Reward structure of cooperation: The decomposed prisoner's dilemma game. *Journal of Personality and Social Psychology, 7*, pp. 21–27.

Pruitt, D. G. (1970). Motivational processes in the decomposed prisoner's dilemma game. *Journal of Personality and Social Psychology, 14*, pp. 227–238.

Selten, R., & Stoecker, R. (1986). End behavior in sequences of finite prisoner's dilemma supergames. *Journal of Economic Behavior and Organization, 7*, pp. 47–70.

Sørensen, R. J., & Grytten, J. (2003). Service production and contract choice in primary physician services. *Health Policy, 66*(1), p. 73.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), pp. 453-458.

Appendix A: Instructions + Comprehension Questions

Welcome to the Experiment!

Preface

You are participating in an economic experiment on decision behavior. You and the other participants will be asked to make decisions for which you can earn money. Your payoff depends on the decisions you make. At the end of the experiment, your payoff will be converted to Euro and paid to you in cash. During the experiment, all amounts are presented in the experimental currency Taler. 10 Taler equals 8 Euro.

The experiment will take about 90 minutes and consists of two parts. You will receive detailed instructions before each part. Note that none of your decisions in either part have any influence on the other part of the experiment.

Part One

Please read the following instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you.

Part one of the experiment consists of 9 rounds of decision situations.

Decision Situations

In each round you take on the role of a physician and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness.

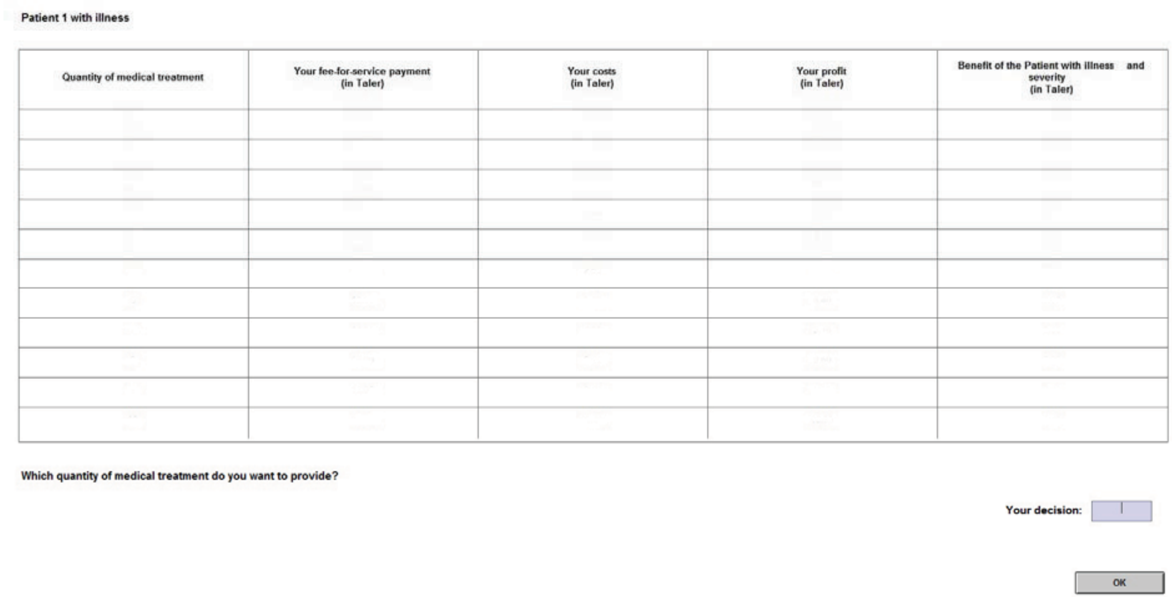
Every patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

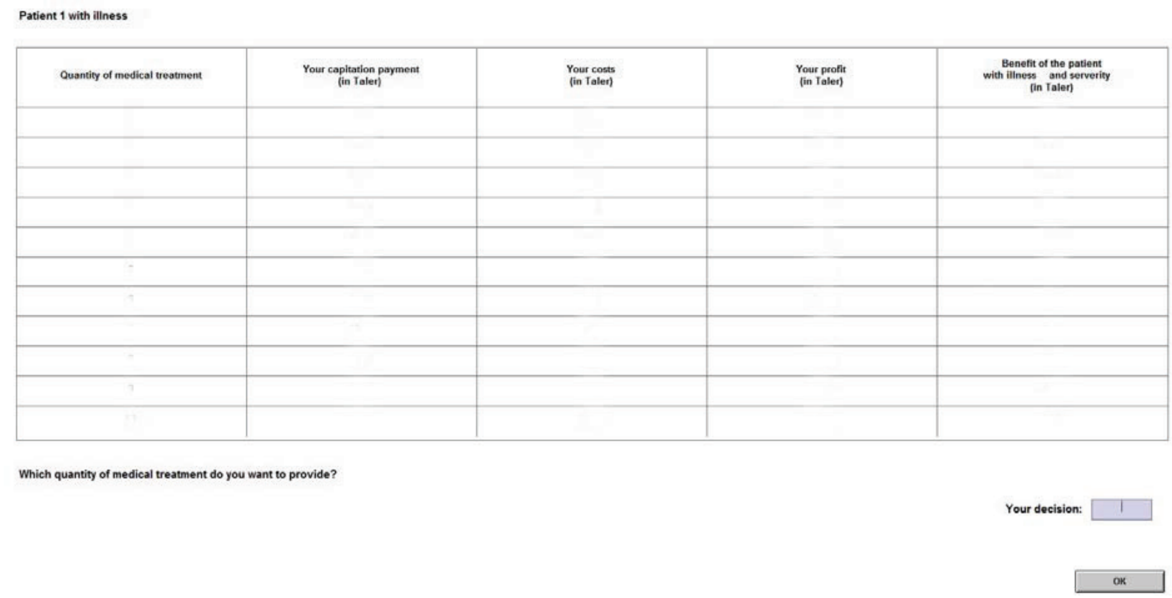
Profit

In each round you receive a fee-for-service (capitation) remuneration for treating the patient. Your remuneration increases with the amount of medical treatment (is irrespective of the amount of medical treatment) you provide. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the fee-for-service (capitation) remuneration.

Every quantity of medical service yields a particular benefit for the patient – contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit but also the patient's benefit.

In each round you will receive detailed information on your screen (see below) for the respective patient the illness, your amount of fee-for-service (capitation) remuneration - for each possible amount of medical treatment - your costs, profit as well as the benefit for the patient with the corresponding illness and severity.





<u>Payment</u>

At the end of the experiment one of the 9 rounds of part one will be chosen at random. Your profit in this round will be paid to you in cash.

For this part of the experiment, no patients are physically present in the laboratory. Yet, the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, an organization which funds the treatment of patients with eye cataract.

The transfer of money to the Christoffel Blindenmission Deutschland e.V. will be carried out after the experiment by the experimenter and one participant. The participant completes a money transfer form, filling in the total patient benefit (in Euro) resulting from the decisions made by all participants in the randomly chosen situation. This form prompts the payment of the designated amount to the Christoffel Blindenmission Deutschland e.V. by the University of Duisburg-Essen's finance department. The form is then sealed in a postpaid envelope and posted in the nearest mailbox by the participant and the experimenter.

After the entire experiment is completed, one participant is chosen at random to oversee the money transfer to the Christoffel Blindenmission Deutschland e.V. The participant receives an additional compensation of 5 Euro for this task. The participant certifies that the process has been completed as described here by signing a statement which can be inspected by all participants at the office of the Chair of Quantitative Economic Policy. A receipt of the bank transfer to the Christoffel Blindenmission Deutschland e.V. may also be viewed here.

Comprehension Questions

Prior to the decision rounds we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part one of the experiment will begin once all participants have answered the comprehension questions correctly.

Part II

Please read the following instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you.

Part two of the experiment also consists of 9 rounds of decision situations.

Decision Situations

As in part one of the experiment, you take on the role of a physician in each round and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness.

Every patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Profit

In each round you are remunerated for treating the patient. In each round you receive a fee-for-service (capitation) remuneration for treating the patient. Your remuneration increases with the amount of medical treatment (is irrespective of the amount of medical treatment) you provide. In addition to this, in each round you receive a capitation remuneration which is irrespective of the amount of medical treatment (a fee-for-service remuneration which increases with the amount of medical treatment). You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the sum of your fee-for-service (capitation) and capitation (fee-for-service) remuneration.

As in part one, every quantity of medical service yields a particular benefit for the patient – contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit but also the patient's benefit.

In each round you will receive detailed information on your screen (see below) for the respective patient the illness, your amount of fee-for-service (capitation) remuneration - for each possible amount of medical treatment - the amount of your capitation (fee-for-service) remuneration, your costs, profit as well as the benefit for the patient with the corresponding illness and severity.

Patient 1 with illness

Quantitity of medical treatment	Your fee-for-service payment (in Taler)	Your capitation payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and severity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

OK

Patient 1 with illness

Quantity of medical treatment	Your capitation payment (in Taler)	Your fee-for-service payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and serverity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

OK

<u>Payment</u>

At the end of the experiment one of the 9 rounds of part two will be chosen at random. Your profit in this round will be paid to you in cash, in addition to your payment from the round chosen for part one of the experiment.

After the experiment is over, please remain seated until the experimenter asks you to step forward. You will receive your payment at the front of the laboratory before exiting the room.

As in part one, no patients are physically present in the laboratory for part two of the experiment. Yet, the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, an organization which funds the treatment of patients with eye cataract.

The process for the transfer of money to the Christoffel Blindenmission Deutschland e.V. as described for part one of the experiment will be carried out by the experimenter and one participant.

Comprehension Questions

Prior to the decision rounds we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part two of the experiment will begin once all participants have answered the comprehension questions correctly.

Finally, we kindly ask you to not talk to anyone about the content of this session in order to prevent influencing other participants after you. Thank you for your collaboration!

Comprehension Questions Part *I*: CAP (FFS)

Questions Tables 1-4:

1-4 a) What is the capitation (fee-for-service)?
1-4 b) What are the costs?
1-4 c) What is the profit?
1-4 d) What is the patient benefit?

Quantity of medical treatment	Capitation (Fee-for-service) (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	20.00 (0.00)	0.00	20.00 (0.00)	15.00
1	20.00 (4.00)	0.20	19.80 (3.80)	16.00
2	20.00 (8.00)	0.80	19.20 (7.20)	17.00
3	20.00 (12.00)	1.80	18.20 (10.20)	18.00
4	20.00 (16.00)	3.20	16.80 (12.80)	19.00
5	20.00 (20.00)	5.00	15.00 (15.00)	20.00
6	20.00 (24.00)	7.20	12.80 (16.80)	19.00
7	20.00 (28.00)	9.80	10.20 (18.20)	18.00
8	20.00 (32.00)	12.80	7.20 (19.20)	17.00
9	20.00 (36.00)	16.20	3.80 (19.80)	16.00
10	20.00 (40.00)	20.00	0.00 (20.00)	15.00

1. Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.
2. Assume that a physician wants to provide 9 quantities of medical treatment for the patient depicted above.

Quantity of medical treatment	Capitation (Fee-for-service) (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness G and severity z (in Taler)
0	20.00 (0.00)	0.00	20.00 (0.00)	10.00
1	20.00 (4.00)	0.20	19.80 (3.80)	12.00
2	20.00 (8.00)	0.80	19.20 (7.20)	14.00
3	20.00 (12.00)	1.80	18.20 (10.20)	16.00
4	20.00 (16.00)	3.20	16.80 (12.80)	18.00
5	20.00 (20.00)	5.00	15.00 (15.00)	20.00
6	20.00 (24.00)	7.20	12.80 (16.80)	22.00
7	20.00 (28.00)	9.80	10.20 (18.20)	24.00
8	20.00 (32.00)	12.80	7.20 (19.20)	22.00
9	20.00 (36.00)	16.20	3.80 (19.80)	20.00
10	20.00 (40.00)	20.00	0.00 (20.00)	18.00

3. Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.
4. Assume that a physician wants to provide 9 quantities of medical treatment for the patient depicted above.

Comprehension Questions Part *II*: Mix-more-CAP(FFS)

1-4 a) What is the capitation (fee-for-service)?
1-4 a) What is the fee-for-service (capitation)?
1-4 b) What are the costs?
1-4 c) What is the profit?
1-4 d) What is the patient benefit?

Quantity of medical treatment	Capitation (Fee-for-service (in Taler)	Fee-for-Service (Capitation) (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	19.20 (0.00)	0.00 (7.20)	0.00	19.20 (7.20)	15.00
1	19.20 (3.20)	0.80 (7.20)	0.20	19.80 (10.20)	16.00
2	19.20 (6.40)	1.60 (7.20)	0.80	20.00 (12.80)	17.00
3	19.20 (9.60)	2.40 (7.20)	1.80	19.80 (15.00)	18.00
4	19.20 (12.80)	3.20 (7.20)	3.20	19.20 (16.80)	19.00
5	19.20 (16.00)	4.00 (7.20)	5.00	18.20 (18.20)	20.00
6	19.20 (19.20)	4.80 (7.20)	7.20	16.80 (19.20)	19.00
7	19.20 (22.40)	5.60 (7.20)	9.80	15.00 (19.80)	18.00
8	19.20 (25.60)	6.40 (7.20)	12.80	12.80 (20.00)	17.00
9	19.20 (28.80)	7.20 (7.20)	16.20	10.20 (19.80)	16.00
10	19.20 (32.00)	8.00 (7.20)	20.00	7.20 (19.20)	15.00

1. Assume that a physician wants to provide 1 quantities of medical treatment for the patient depicted above.
2. Assume that a physician wants to provide 8 quantities of medical treatment for the patient depicted above.

Quantity of medical treatment	Capitation (Fee-for-service (in Taler)	Fee-for-Service (Capitation) (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness G and severity z (in Taler)
0	19.20 (0.00)	0.00 (7.20)	0.00	19.20 (7.20)	10.00
1	19.20 (3.20)	0.80 (7.20)	0.20	19.80 (10.20)	12.00
2	19.20 (6.40)	1.60 (7.20)	0.80	20.00 (12.80)	14.00
3	19.20 (9.60)	2.40 (7.20)	1.80	19.80 (15.00)	16.00
4	19.20 (12.80)	3.20 (7.20)	3.20	19.20 (16.80)	18.00
5	19.20 (16.00)	4.00 (7.20)	5.00	18.20 (18.20)	20.00
6	19.20 (19.20)	4.80 (7.20)	7.20	16.80 (19.20)	22.00
7	19.20 (22.40)	5.60 (7.20)	9.80	15.00 (19.80)	24.00
8	19.20 (25.60)	6.40 (7.20)	12.80	12.80 (20.00)	22.00
9	19.20 (28.80)	7.20 (7.20)	16.20	10.20 (19.80)	20.00
10	19.20 (32.00)	8.00 (7.20)	20.00	7.20 (19.20)	18.00

3. Assume that a physician wants to provide 1 quantities of medical treatment for the patient depicted above.
4. Assume that a physician wants to provide 8 quantities of medical treatment for the patient depicted above.

Appendix B: Parameter Tables

Table B.1

Treatment	Variable	Quantity (q) 0	1	2	3	4	5	6	7	8	9	10
A-C2	R_{kl}^{I}	10	10	10	10	10	10	10	10	10	10	10
	$R_{kl}^{II\ CAP}$	9.6	9.6	9.6	9.6	9.6	9.6	9.6	9.6	9.6	9.6	9.6
	$R_{kl}^{II\ FFS}$	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2	3.6	4
A-C4	R_{kl}^{I}	10	10	10	10	10	10	10	10	10	10	10
	$R_{kl}^{II\ CAP}$	8.4	8.4	8.4	8.4	8.4	8.4	8.4	8.4	8.4	8.4	8.4
	$R_{kl}^{II\ FFS}$	0	0.8	1.6	2.4	3.2	4	4.8	5.6	6.4	7.2	8
A-F8	R_{kl}^{I}	0	2	4	6	8	10	12	14	16	18	20
	$R_{kl}^{II\ CAP}$	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6
	$R_{kl}^{II\ FFS}$	0	1.6	3.2	4.8	6.4	8	9.6	11.2	12.8	14.4	16
A-F6	R_{kl}^{I}	0	2	4	6	8	10	12	14	16	18	20
	$R_{kl}^{II\ CAP}$	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.4
	$R_{kl}^{II\ FFS}$	0	1.2	2.4	3.6	4.8	6	7.2	8.4	9.6	10.8	12
NA-C2	R_{kl}^{I}	10	10	10	10	10	10	10	10	10	10	10
	$R_{kl}^{II\ CAP}$	8	8	8	8	8	8	8	8	8	8	8
	$R_{kl}^{II\ FFS}$	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2	3.6	4
NA-CF8	R_{kl}^{I}	0	2	4	6	8	10	12	14	16	18	20
	$R_{kl}^{II\ CAP}$	2	2	2	2	2	2	2	2	2	2	2
	$R_{kl}^{II\ FFS}$	0	1.6	3.2	4.8	6.4	8	9.6	11.2	12.8	14.4	16
all	c_{kl}	0	0.1	0.4	0.9	1.6	2.5	3.6	4.9	6.4	8.1	10
A-C2	π_{kl}^{I}	10	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6	1.9	0
	π_{kl}^{II}	9.6	9.9	10	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6
A-C4	π_{kl}^{I}	10	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6	1.9	0
	π_{kl}^{II}	8.4	9.1	9.6	9.9	10	9.9	9.6	9.1	8.4	7.5	6.4
A-F8	π_{kl}^{I}	0	1.9	3.6	5.1	6.4	7.5	8.4	9.1	9.6	9.9	10
	π_{kl}^{II}	3.6	5.1	6.4	7.5	8.4	9.1	9.6	9.9	10	9.9	9.6
A-F6	π_{kl}^{I}	0	1.9	3.6	5.1	6.4	7.5	8.4	9.1	9.6	9.9	10
	π_{kl}^{II}	6.4	7.5	8.4	9.1	9.6	9.9	10	9.9	9.6	9.1	8.4
NA-C2	π_{kl}^{I}	10	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6	1.9	0
	π_{kl}^{II}	8	8.3	8.4	8.3	8	7.5	6.8	5.9	4.8	3.5	2
NA-CF8	π_{kl}^{I}	0	1.9	3.6	5.1	6.4	7.5	8.4	9.1	9.6	9.9	10
	π_{kl}^{II}	2	3.5	4.8	5.9	6.8	7.5	8	8.3	8.4	8.3	8
all	B_{Ax}	4	5	6	7	6	5	4	3	2	1	0
	B_{Ay}	2	3	4	5	6	7	6	5	4	3	2
	B_{Az}	0	1	2	3	4	5	6	7	6	5	4
	B_{Bx}	7	8	9	10	9	8	7	6	5	4	3
	B_{By}	5	6	7	8	9	10	9	8	7	6	5
	B_{Bz}	3	4	5	6	7	8	9	10	9	8	7
	B_{Cx}	8	10	12	14	12	10	8	6	4	2	0
	B_{Cy}	4	6	8	10	12	14	12	10	8	6	4
	B_{Cz}	0	2	4	6	8	10	12	14	12	10	8

Table B.2 Experimental Parameters in Condition P-C2

Variable	0	1	2	3	4	5	6	7	8	9	10
R_{kl}^{I}	15	15	15	15	15	15	15	15	15	15	15
$R_{kl}^{II\ CAP}$	8	8	8	8	8	8	8	8	8	8	8
$R_{kl}^{II\ FFS}$	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2	3.6	4
$c_{kl}^{I\ Pres}$	7	6.7	6.6	6.7	7	7.5	8.2	9.1	10.2	11.5	13
c_{kl}^{II}	0	0.1	0.4	0.9	1.6	2.5	3.6	4.9	6.4	8.1	10
$\pi_{kl}^{I\ Pres}$	8	8.3	8.4	8.3	8	7.5	6.8	5.9	4.8	3.5	2
π_{kl}^{II}	8	8.3	8.4	8.3	8	7.5	6.8	5.9	4.8	3.5	2

Chapter 3

How Effective are Pay-for-Performance Incentives for Physicians? A Laboratory Experiment

Reference

Brosig-Koch, J., Hennig-Schmidt, H., Kairies, N., and D. Wiesen. 2013b. "How Effective are Pay-for-Performance Incentives for Physicians? A Laboratory Experiment." *Ruhr Economic Papers* No. 413.

RUHR
ECONOMIC PAPERS

Jeannette Brosig-Koch
Heike Hennig-Schmidt
Nadja Kairies
Daniel Wiesen

How Effective are Pay-for-Performance Incentives for Physicians?

A Laboratory Experiment

UNIVERSITÄT DUISBURG ESSEN

#413

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Editorial Office

Ruhr Economic Papers #413

Responsible Editor: Volker Clausen

ISSN 1864-4872 (online) – ISBN 978-3-86788-468-6

Ruhr Economic Papers #413

Jeannette Brosig-Koch, Heike Hennig-Schmidt,
Nadja Kairies, and Daniel Wiesen

How Effective are Pay-for-Performance Incentives for Physicians?

A Laboratory Experiment

UNIVERSITÄT DUISBURG ESSEN

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

http://dx.doi.org/10.4419/86788468
ISSN 1864-4872 (online)
ISBN 978-3-86788-468-6

Jeannette Brosig-Koch, Heike Hennig-Schmidt, Nadja Kairies, and Daniel Wiesen[1]

How Effective are Pay-for-Performance Incentives for Physicians? – A Laboratory Experiment

Abstract

Recent reforms in health care have introduced a variety of pay-for-performance programs using financial incentives for physicians to improve the quality of care. Their effectiveness is, however, ambiguous as it is often difficult to disentangle the effect of financial incentives from the ones of various other simultaneous changes in the system. In this study we investigate the effects of introducing financial pay-for-performance incentives with the help of controlled laboratory experiments. In particular, we use fee-for-service and capitation as baseline payment schemes and test how additional pay-for-performance incentives affect the medical treatment of different patient types. Our results reveal that, on average, patients significantly benefit from introducing pay-for-performance, independently of hether it is combined with capitation or fee-for-service incentives. The magnitude of this effect is significantly influenced by the patient type, though. These results hold for medical and non-medical students. A cost-benefit analysis further demonstrates that, overall, the increase in patient benefits cannot overcompensate the additional costs associated with pay-for-performance. Moreover, our analysis of individual data reveals different types of responses to pay-for-performance incentives. We find some indication that pay-forperformance might crowd out the intrinsic motivation to care for patients. These insights help to understand the effects caused by introducing pay-for-performance schemes.

JEL Classification: C91, I11

Keywords: Physician incentive schemes; pay-for-performance; fee-for-service; capitation; laboratory experiment

April 2013

1 Jeannette Brosig-Koch, Nadja Kairies, Daniel Wiesen, University of Duisburg-Essen; Heike Hennig-Schmidt, University of Bonn. *– We thank the participants of the European Conference on Health Economics in Zurich, of the Economic Science Association International Conference in New York and the Economic Science Association European Conference in Cologne for their valuable comments and suggestions. Financial support provided by the Deutsche Forschungsgemeinschaft (German Research Foundation) is gratefully acknowledged.* *– All correspondence to Jeannette Brosig-Koch, University of Duisburg-Essen, Faculty of Economics and Business Administration, Universitätsstr. 12, 45117 Essen, Germany, E-Mail: jeannette.brosig-koch@ibes.uni-due.de.*

1. Introduction

For policy makers, improving quality in the health care market has been a major issue in recent years (McCellan, 2011). Thus, the reform of reimbursement for health care providers, i.e. physicians, has become seminal. As established capitation or fee-for-service schemes rather "pay" physicians for the quantity than for the quality of medical treatment, policy makers have started to implement monetary *pay-for-performance* (P4P) schemes to incentivize quality improvement. By now, payment schemes including P4P incentives have been introduced in many countries, e.g. the UK (Doran et al., 2006, Campbell et al., 2009), the USA (Rosenthal et al., 2004, and Rosenthal, 2008), Australia (Duckett et al. 2008, Scott 2008, Scott et al. 2009), France (Mousquès et al., 2012), Korea (Kim, 2010), New Zealand (Perkins and Seddon, 2006, and Buetow, 2008), or Spain (Gené-Badia et al. , 2007)**.**

Existing P4P incentives vary to some extent, the key determinants are alike, however. First, P4P incentives are usually designed as a monetary bonus paid in addition to the basic payment, which usually is capitation (CAP), fee-for-service (FFS), or a combination of both.[1] P4P programs with a basic payment of CAP have been introduced, among others, in the UK within the Quality and Outcomes Framework (QOF), in the US within the California P4P Program, and in Spain. Countries that have implemented P4P programs with a basic FFS payment scheme include, e.g., the US (Medicare), Australia, New Zealand and France. Second, performance measures usually include a combination of structure, process and outcome measures (Donabedian, 2005). Third, the basis for the financial bonus is either an absolute measure such as targets or intervals (see, e.g., the QOF in the UK), based on improvements in the measure (see, e.g., the Medicare Physician Group Practice Demonstration in the US), or a relative ranking (see, e.g., the Value Incentive Program in Korea). Fourth, the size of the bonus is rather small, i.e., less than 5 percent of the entire remuneration (Borowitz et al., 2010). An exception is the QOF in the UK where the bonus amounts to about 20 percent (Doran et al., 2006).

Despite the increasing popularity of monetary P4P, it is yet unclear whether they actually enhance quality and efficiency of care (Borowitz et al., 2010). Reasons for this uncertainty include limited access to valid data and design problems leading to negative effects (see Maynard, 2012, for a survey). The latter comprise, among others, substituting away from the non-rewarded towards the rewarded aspects of quality (Mullen et al., 2010), improving quality to the performance target (Campbell et al., 2009), or gaming with quality indicators in terms of exception reporting (Gravelle et al., 2010). For the California P4P Program, Mullen et al. (2010), e.g., find that P4P did have a positive impact on some of the rewarded clinical measures. However, in the cases in which physicians substituted away from unrewarded aspects of medical treatment, the gain in quality of the P4P indicator could sometimes not be offset by the reduction in the other indicators. In the UK, incentives within

[1] Most P4P programs include incentives for primary care physicians, specialists, or hospitals. Since we will focus on the direct implications of introducing P4P incentives on physicians' behavior, hospitals are not considered as for their case P4P payments imply additional agency problems. Accordingly, specific payment forms used in hospitals such as budgets or DRGs are also not discussed in this study.

the QOF seemed to have resulted only in a short-term increase in quality due to quality thresholds (Campbell et al., 2009) and even triggered unintended gaming with the quality indicators, e.g., in terms of exception reporting (Gravelle et al., 2010).

While these examples underline problems of existing P4P schemes resulting from the specific designs of performance measures (in particular the difficulty of identifying quality), they do not tell much about the effectiveness of P4P incentives *per se*. Moreover, field research on P4P programs often faces the problem that relevant parameters are difficult to control like the implementation of additional incentives (e.g., public reporting), regional or institutional characteristics of study groups, or the health status of patients treated.[2] Hence, it is difficult to assess whether the quality metrics used are not reliable or whether the financial P4P incentives themselves fail to work. That introducing monetary rewards can have negative behavioral effects has been demonstrated in previous field and laboratory research. For example, running an IQ task and a donation collection task Gneezy and Rustichini (2000) observe that implementing a monetary bonus can significantly decrease work effort. Similar results have been obtained by Frey and Oberholzer-Gee (1997) on the willingness to host a nuclear waste repository and by Mellström and Johannesson (2008) on the supply of blood donors (see also Frey and Jegen, 2001, for a survey). Accordingly, it is an open question whether introducing P4P incentives aimed at achieving an optimal provision of medical care actually crowds out physicians' intrinsic motivation.

In order to prevent costly failures and to guarantee the success of P4P programs, a controlled analysis of the effects of financial P4P incentives on physician provision behavior is of great importance. The purpose of our paper is to provide such a controlled analysis of P4P incentives. In particular, our study contributes to previous research by investigating the effects of P4P incentives on the provision of medical care in a controlled laboratory experiment. Laboratory experiments are a new and emerging method in health economics (see, e.g., Hennig-Schmidt et al., 2011). It allows analyzing individual behavior in a controlled environment where only one parameter is changed at a time. External factors like patients' health status can be isolated and, if a change in behavior occurs, it can be attributed to the change of that one parameter, here the change of the payment scheme. Moreover, in the laboratory we can employ a reliable measure for patients' health status and, accordingly, the quality of care making it possible to isolate the effect of monetary P4P incentives from effects of deficient quality indicators. Laboratory experiments are a relatively inexpensive method to study behavioral responses to reforms. They are, thus, an effective additional tool to guide policy makers in designing appropriate incentive schemes.

In our experiment, participants act in the role of physicians and make decisions about the medical treatment of nine different patients. The treatment choice affects the patient benefit. The patients differ systematically in their illness and the degree of severity of these illnesses. We induce a sequential within-subject design to account for a payment reform given the same patient population before and after implementation. In part one of the

[2] Many of the P4P reimbursement schemes like the Quality and Outcomes Framework in the UK or Pacificare in California have been implemented along with public evaluation programs.

experiment, physicians decide about the quantity of medical treatment either under a capitation or a fee-for-service payment scheme. In part two, they are confronted with a P4P scheme based on capitation and fee-for-service payment, respectively. P4P is designed as a performance-based bonus paid in case physicians treat patients close to their optimum. The bonus applies to all patients; hence substitution away from unrewarded patients or exception reporting is impossible. The experimental design not only allows a controlled investigation of physicians' responses to P4P incentives, but also enables us to compare cost-benefit ratios between the different payment schemes. Moreover, due to our within-subject design we are able to identify whether and how patterns of behavior revealed in the basic payment schemes change with P4P incentives.

Our results demonstrate that introducing P4P incentives leads to treatment levels which are significantly closer to patient optimal levels than those observed with the baseline incentives of CAP and FFS. The degree to which patients benefit from P4P depends on their individual characteristics (i.e., the severity of their illness). These results hold for both, medical and non-medical student participants. Moreover, comparing patient benefits and physician remuneration between the different payment schemes reveals that the increase in benefits is outweighed by the additional cost of P4P incentives. Analyzing individual treatment behavior shows that there are different behavioral types and that P4P incentives differently affect these types. In particular, there is some indication that the financial incentives included in P4P might crowd out the intrinsic motivation to care.

The paper is organized as follows. In Section 2, we present design and procedure of our experiment with a special focus on how we designed the P4P incentives. In Section 3 we present our results and in Section 4 we conclude.

2. Experimental Design and Procedure

In the experiment, we analyze the impact of financial P4P incentives on the quantity of medical treatment and, thus, the patient health benefit. Capturing the character of a reform we use a sequential design consisting of two parts. In part 1 of the experiment, subjects are confronted with either one of the baseline payment schemes, CAP or FFS. In part 2, the only parameter we change is introducing a P4P incentive. This procedure allows a controlled ceteris paribus analysis.

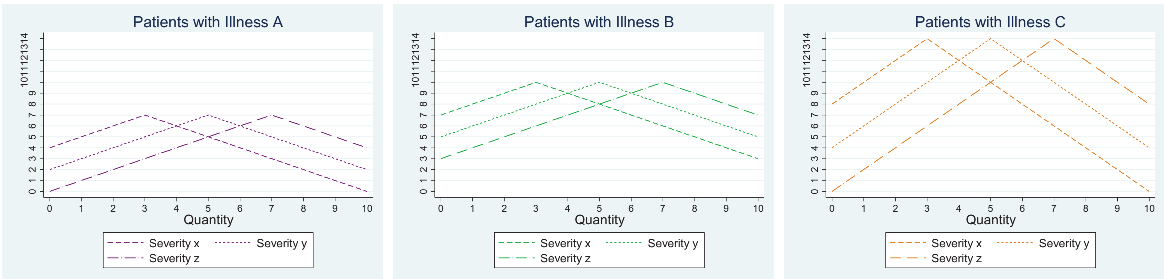
The basic design of the **decision situation** follows Brosig-Koch et al. (2013). Subjects are in the role of a physician i and decide on the quantity of medical services $q = \{0, 1, \dots, 10\}$ for nine different patients $j \in [0,1,\dots,9]$ who vary in their illnesses $k \in [A, B, C]$, and the severities $l \in [x, y, z]$ of these illnesses. For each patient a physician receives a remuneration R depending on the experimental condition and bears costs $c_{kl}(q) = 0.1 \cdot q^2$ which are the same in all conditions.[3] With each decision, the physician determines her profit π_{kl}^{i} , i.e. $R - c_{kl}(q)$, as well as the patient's health benefit B_{kl}. Both, profits and patient benefits, are

[3] Convex cost functions are used in several theoretical models describing physician behavior (e.g., Ma, 1994 and Choné and Ma, 2011).

measured in monetary terms. Patients in the experiment are assumed to be passive and fully insured, accepting the quantity of medical treatment provided by the physician.

Patient benefit functions (see Figure 1) take on a global optimum on the quantity interval [0,10].[4] The optimal level of patient benefit depends on the illnesses, i.e. $B_{Al}(q^*) = 7$, $B_{Bl}(q^*) = 10$, and $B_{Cl}(q^*) = 14$, and is achieved when the patient optimal quantity q^* is provided. Moreover, each illness implies a certain slope of the benefit function, i.e., a certain change of benefit resulting from an additional unit of medical service. While the slope of the benefit function is the same for illnesses A and B, it is different for illness C. The patient optimal quantity q^* depends on the severity of an illness, i.e. $q^* = 3$ for severity *x*, $q^* = 5$ for severity *y*, and $q^* = 7$ for severity *z*. Taking q^* as a benchmark, we can identify the magnitude of overprovision and underprovision. The optimal amount of medical services is specified for each patient and is known to the physician.[5]

Figure 1: Patient Benefit Functions for Illnesses k and Severities l



Patients in the experiment are not present in the lab. The experimental participants are informed, however, that the monetary equivalent to the realized patient benefit is transferred to a charity (the *Christoffel Blindenmission*) caring for real patients with eye cataract (see also section 2.2). In the following, we explain the payment schemes used in the experimental conditions in more detail.

2.1 Experimental Conditions

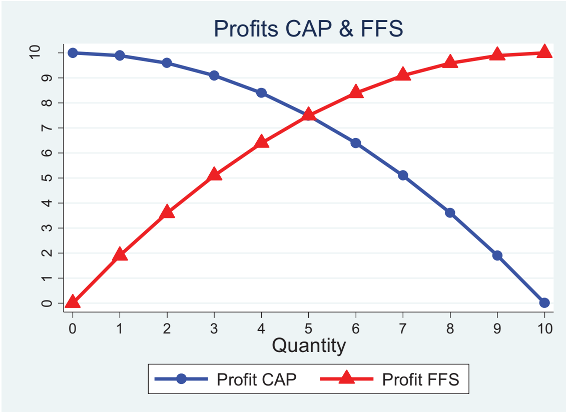
2.1.1. Baseline Incentive Schemes (Part 1)

The design of the baseline incentives CAP and FFS follows Brosig-Koch et al. (2013). Under CAP, a physician receives a lump-sum payment per patient, i.e. $R = LS = 10$. Thus, physician i's profit per patient is $\pi^i_{kl}(q) = 10 - c_{kl}(q)$. Under FFS, physicians are paid a fee of $p = 2$ per service they provide, i.e. $R(q) = 2q$. Accordingly, physician i's profit per patient is $\pi^i_{kl}(q) = 2q - c_{kl}(q)$ (see Figure 2).

[4] A concave patient benefit function has been widely assumed in theoretical papers; see e.g., Ellis and McGuire, 1986, Ma, 1994, Choné and Ma, 2011.

[5] Thus, there is no uncertainty about the impact of the chosen quantity of medical services on the patient's health benefit.

Figure 2: Profits of CAP and FFS Schemes in Part1 of the Experiment



The maximum profit a physician can achieve is identical for CAP and FFS ($\hat{\pi}_{kl}^{CAP} = \hat{\pi}_{kl}^{FFS} = 10$), as is the (absolute value of) marginal changes of profits. The profit maximizing quantity of medical services $\hat{q}$, however, differs and is 0 for CAP and 10 for FFS (see also Appendix A.3 for the complete set of parameter values). Given this parameterization, the profit functions for the two payment schemes are perfectly symmetric. Since patient benefit functions are also symmetric, we are able to fully compare behavior revealed under the two payment schemes.

2.1.2. P4P Incentive Scheme (Part 2)

P4P programs for primary care physicians and specialists use financial incentives to stimulate improvements in the quality of care and, in some cases, reductions in costs. Motivated by the composition of actual P4P schemes we chose our experimental parameters. We use the CAP and FFS payment schemes as a basis and provide an additional performance-based bonus in case the physician provides treatment within a predetermined performance interval. We link the performance measure to the individual patient's health benefit that can be interpreted as the health outcome of a certain medical procedure.

In particular, we design the P4P scheme as follows. P4P incentives apply to all patients; hence we exclude problems of multitasking or exception reporting. Physicians are rewarded for being "close" to the patient optimal quantity of medical care q^*, which is perfectly identifiable in our design. A physician is rewarded with a bonus for those quantity choices that do not differ by more than *one* unit from the patient optimal quantity q^*. Whenever the physician's quantity choice differs by more than one unit, she does not receive a bonus. The bonus is specified such that we provide higher incentives to treat those patients optimally that are in need of a high (low) quantity of medical services in CAP (FFS). More specifically, we designed the bonus in a way that maximum incentives are fully comparable across different patient types. This means, physicians receive the same maximum profit when treating the patients within the predetermined performance interval, irrespective of the severity of the illness.[6] We increased the physicians' profit maximum by 20 percent to $\hat{\pi}$=12 for all patient types to achieve a sufficient difference in incentives to the baseline schemes. This increase is in line with the Quality and Outcomes Framework in the UK which was intended to increase family physicians' income up to 25 percent depending on their performance (Doran et al., 2006). The remuneration from P4P is

$$\text{P4P} = \begin{cases} P_l & if\ |q - q^*| \leq 1 \\ 0 & otherwise \end{cases}$$

where P_l is the P4P bonus which depends on the patient's severity of illness *l*. Given that the physician's profit function in CAP is symmetric to the profit function in FFS, and given that the maximum profit with P4P incentives is equal to 12 for all patient types, we have symmetric profit functions also with P4P incentives. Table 1 shows the bonus for each severity in CAP and in FFS.

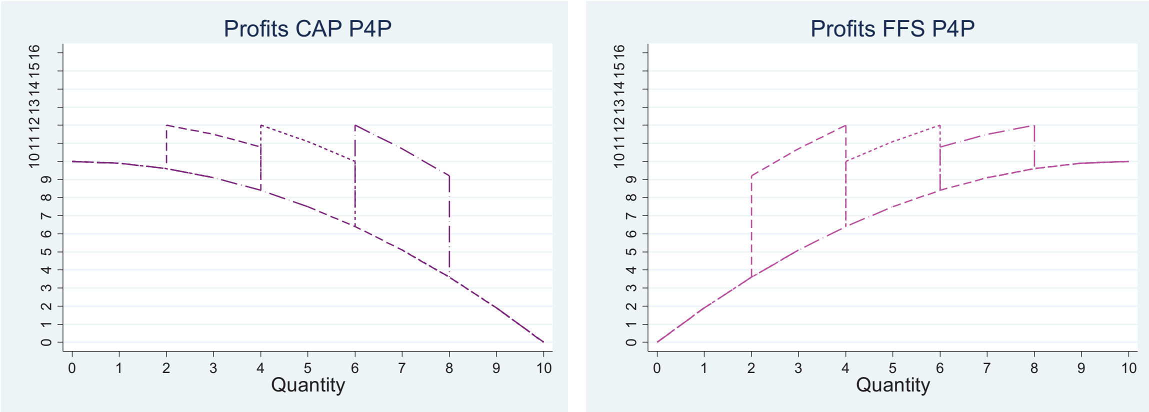
Table 1: Parameters for the P4P bonus in CAP and FFS

Severity	P_l^{CAP}	P_l^{FFS}
x (*q**=3)	2.4	5.6
y (*q**=5)	3.6	3.6
z (*q**=7)	5.6	2.4

Physicians' total profits under the P4P payment schemes are calculated by $\pi_{kl}^{i} = R_{kl}^{CAP/FFS} P_k^{CAP/FFS} - 0.1q^2$ and are depicted in Figure 3 (see also Appendix A.3 for the complete set of parameter values). By paying the bonus if the chosen quantity is within an interval of quantities instead of if it exactly matches the patient optimal quantity, the physician's and the patient's interests are not perfectly aligned. The physician profit maximizing quantities are $\hat{q}_x^{CAP\ P4P}$=2, $\hat{q}_y^{CAP\ P4P}$=4, $\hat{q}_z^{CAP\ P4P}$=6 under CAP P4P and $\hat{q}_x^{FFS\ P4P}$=4, $\hat{q}_y^{FFS\ P4P}$=6, $\hat{q}_z^{FFS\ P4P}$=8 under FFS P4P, whereas the patient optimal quantities are q_x^*=3, q_y^*= 5 and q_z^*=7. Hence, the incentive from the baseline schemes are still inherent in the P4P schemes, but to a substantially lower extent.

[6] Recall that the severity of an illness determines the patient optimal quantity q*, which is related to a certain payment from the baseline schemes.

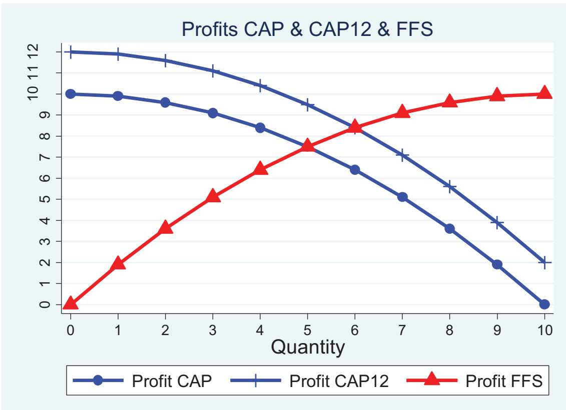
Figure 3: Profits of CAP P4P and FFS P4P Schemes in Part 2 of the Experiment



2.1.3. Constant Profit Maximum

Policy-makers who aim at reforming physician payment towards a P4P scheme are often constrained to a constant budget. Assuming selfish physicians, it appears reasonable to keep the profit maximum constant before and after reforming a payment scheme. To test the effect of a constant profit maximum, we model a CAP payment scheme with a profit maximum equal to the one in CAP P4P $\hat{\pi}$=12, i.e. CAP12. The lump-sum payment is assumed to be 12 resulting in physician *i*'s profit $\pi_{kl}^{i}(q) = 12 - c_{kl}(q)$ (see Figure 4).

Figure 4: Profits of CAP, FFS and CAP12 Schemes



2.2. Experimental Protocol

The computerized experiment was programmed with z-Tree (Fischbacher, 2007) and conducted at elfe, the Essen Laboratory for Experimental Economics of the University of Duisburg-Essen, Germany. Overall 44 medical students (who are supposed to become physicians in the future) and 56 non-medical students from the University of Duisburg-Essen participated in our experimental sessions. See Table 2 for an overview of experimental conditions. Subjects were recruited by the online recruiting system ORSEE (Greiner, 2004).[7]

Table 2: Number of Subjects per Condition and Type of Students

Condition	Payment Scheme Part 1 - Part 2	Type of Students		**Total**
		Non-Medical	Medical	
C1	CAP - CAP P4P	23	22	**45**
C2	FFS - FFS P4P	20	22	**44**
CAP12	CAP12 - CAP P4P	13	-	**13**
Total		**56**	**44**	**100**

The procedure follows Brosig-Koch et al. (2013). Prior to the actual experiment subjects were randomly assigned to their cubicles. They had enough time for reading the instructions for part 1 of the experiment (when deciding under one of the baseline payment schemes) and for privately asking the experimenter clarifying questions. After reading the instructions, subjects had to answer several control questions to make sure they had understood the decision task. Once everyone had answered these questions correctly, part 1 of the experiment started. The order of the nine patients to be treated was randomly determined and kept constant for all subjects in part 1 and part 2 of all conditions to make the data straightforwardly comparable across payment schemes (see Table 3).

Table 3: Randomized Order of Illnesses and Severities of Illness

Patient *j*	1	2	3	4	5	6	7	8	9
Illness *k*	B	C	A	B	B	A	C	A	C
Severity *l*	*x*	*x*	*z*	*y*	*z*	*y*	*z*	*x*	*y*

For each of the nine patients, subjects are informed on their decision screen about their remuneration, their cost and profit, as well as about the patient's benefit for each quantity of medical treatment. All monetary amounts are given in the experimental currency Taler the exchange rate being 1 Taler = €0.08.[8] In part 2 of the experiment, we applied exactly the same procedure except for subjects now being confronted with a P4P payment scheme. At the end of part 2, one decision for each part of the experiment was randomly chosen to be relevant for the subject's actual payoff and the patient benefit. We used this procedure to

[7] Students who registered in ORSEE to participate in laboratory experiments at the Essen Laboratory for Experimental Economics were invited via automatically generated e-mails and registered for a special session. We can thus say that subjects were randomly allocated to the experimental conditions. Moreover, subjects were not informed about the content of the experimental conditions unless they participated in a session.

[8] Instructions, control questions, and an example of the decision screen are included in Appendix A.2.

avoid wealth and averaging effects.[9] After the experiment, subjects privately received their payment and were dismissed.

The monetary value of patient benefits for the two payoff-relevant decisions aggregated over all subjects was transferred to the *Christoffel Blindenmission*. To verify this transfer, we applied a procedure similar to the one used in Hennig-Schmidt et al. (2011) and Eckel and Grossman (1996). After the experiment, a previously randomly determined subject acted as our monitor and verified that a correct transfer order on the aggregated benefit in the respective session was written to the university's financial department. The monitor and experimenter then walked together to the nearest mailbox and deposited the order in a sealed envelope. The monitor was paid an additional 5€.

Sessions lasted for about 70 minutes. Subjects earned, on average, 16.58€. The average benefit per patient was 13.38€. In total, 1,351.78€ were transferred to the *Christoffel Blindenmission*. The money supported surgical treatments of cataract patients in a hospital in Masvingo (Zimbabwe) staffed by ophthalmologists of the *Christoffel Blindenmission*. Average costs for such an operation amounted to about 30€. Thus, the money from our experiment allowed treating 45 patients.[10]

3. Results

In this section, we first present the average quantity of treatment physicians provide for each decision under the baseline payment schemes CAP and FFS. Second, we analyze whether patients actually benefit from P4P payment incentives and present a within-subject comparison of treatment behavior before and after a reform towards P4P payment. For this, we restrict the analysis to non-medical students. Third, we compare provision behavior between non-medical students and prospective physicians. Fourth, we check whether our results hold when keeping the profit maximum constant. Finally, we classify subjects according to their behavior in part 1 and describe type-specific changes of behavior in part 2.

3.1. Provision Behavior under Baseline Payment Schemes

Analyzing physician provision behavior under the baseline payment schemes, we find that subjects are not purely selfish in the sense that many of them do not provide their profit-maximizing quantity of treatment $\hat{q}_{kl}^{CAP}$=0 or $\hat{q}_{kl}^{FFS}$=10. However, in CAP patients are significantly underserved, i.e. the average quantities provided per subject per severity are significantly lower than the patient optimal quantities (p=0.0000 Fisher Pitman Permutation test for paired replicates, two-sided; FPPP in the following). In FFS, patients are significantly overserved, i.e. the average quantities provided per subject and severity are significantly higher than the patient optimum (p=0.0000, FPPP two-sided).

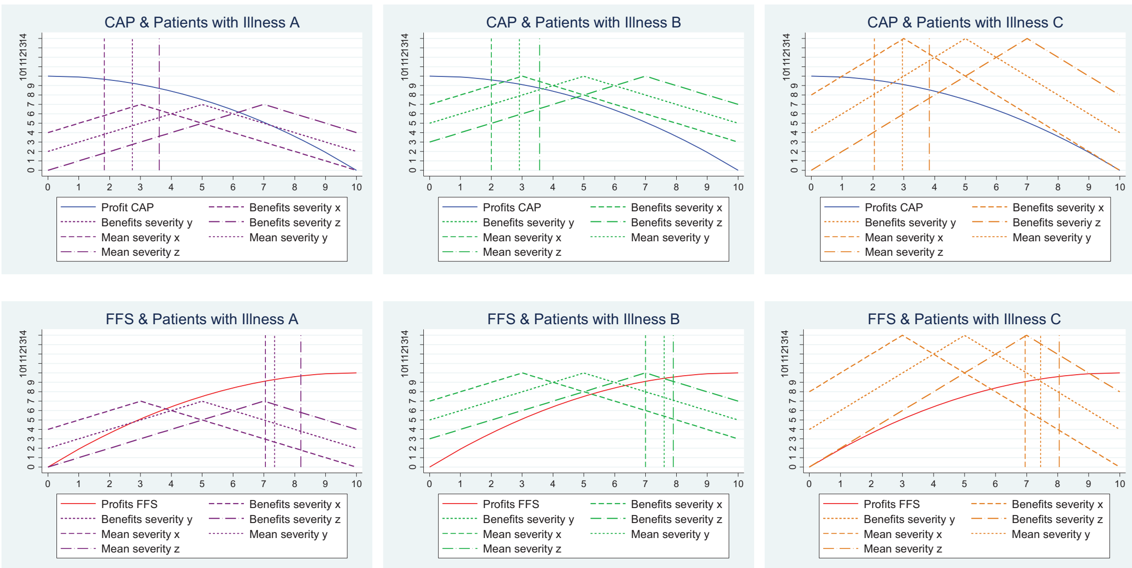
[9] Various studies confirm that the random payment technique does not affect the power of the monetary incentive for non-complex choice tasks (Starmer and Sugden, 1991, Cubitt et al., 1998, Laury, 2006, Baltussen et al., 2012).

[10] Subjects were not informed that the money would be transferred to a developing country in order to avoid motives like compassion.

The symmetry of physicians' profit functions and patients' benefit functions allows comparing the extent of deviations from the patient optimum between CAP and FFS. From a psychological point of view, people might regard choosing a higher quantity of medical treatment for the patient (which is incentivized in FFS) less severe than not providing this quantity to the patient (which is incentivized in CAP). We do not find such an effect, however. In both payment schemes, we observe a similar degree of over- or underprovision, respectively (p=0.5911, Fisher Pitman Permutation test for independent replicates, two-sided; FPPI in the following).

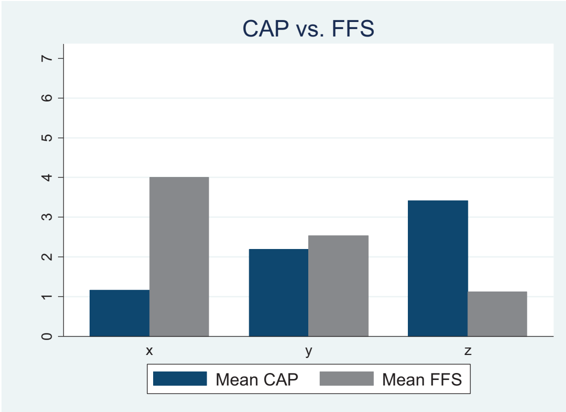
Figure 5 shows the average quantities chosen per patient for each severity (depicted by the vertical lines) as well as physician's profit and patient's benefit for each quantity of medical treatment in CAP and FFS.

Figure 5: Average Treatment Quantity per Severity in CAP and FFS



Testing the influence of patient characteristics on provision behavior yields a significant and systematic effect of the severity of illness (p<0.0746, FPPP, two-sided), but almost no significant effect of the illness itself (p>0.2722, except for 1 out of the 6 comparisons where p=0.0501, FPPP, two-sided). Thus, neither the maximum level of patient health benefit nor the change of benefit that is associated with an additional unit of medical treatment (both are implied by an illness) systematically affect provision behavior. Only the quantity yielding the maximum health benefit (which is implied by the severity of an illness) influences the choice of medical treatment. The more the patient optimal quantity deviates from the physician's profit maximizing quantity (i.e., the more severe the illness is in CAP and the less severe the illness is in FFS), the more does the quantity choice deviate from the patient optimum (p<0.0004, FPPP, two-sided), see Figure 6. These results are fully in line with the observations made by Brosig-Koch et al. (2013).

Figure 6: Distance between Patient Optimum and Chosen Quantity per Severity



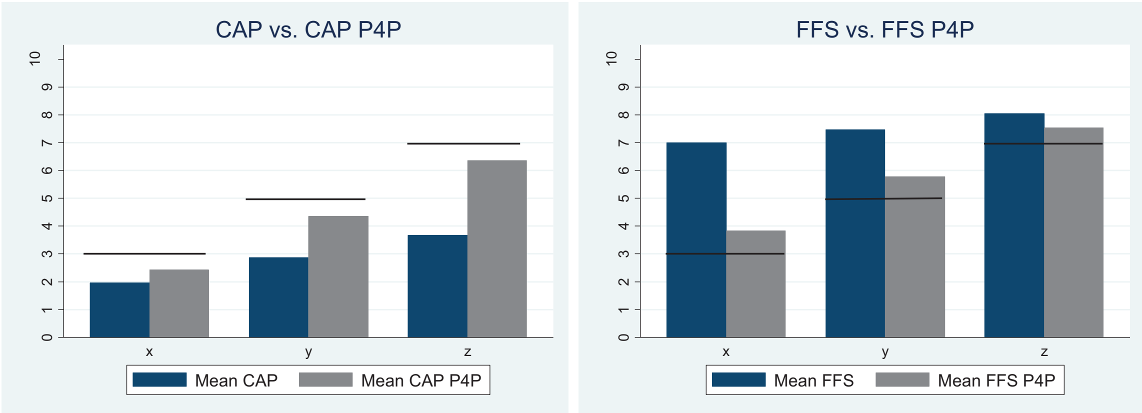
Result 1: Under CAP (FFS) physicians significantly underserve (overserve) patients. Deviations from the patient optimal medical treatment are significantly influenced by the severity of a patient's illness.

3.2. (Change of) Provision Behavior under P4P Incentives

Next, we compare subjects' provision behavior before and after introducing P4P incentives. As a patient's illness has almost no significant (and no systematic) effect on the quantity of treatment provided under P4P incentives ($p>0.6234$, except for 2 out of the 6 comparisons where $p<0.0232$, FPPP, two-sided), we pool decisions over illnesses in all subsequent analyses.

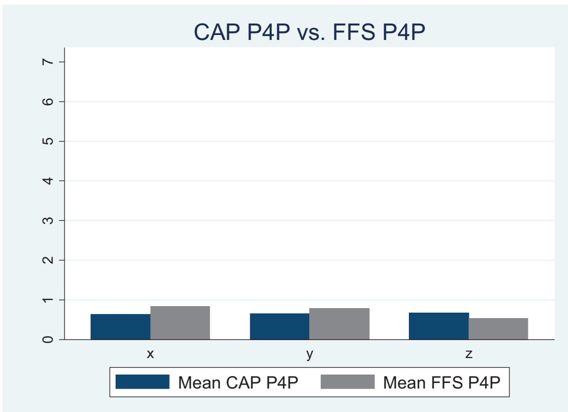
Figure 7 depicts the average treatment quantities for each severity without and with P4P incentives. The black horizontal lines indicate the patient optimal quantities. We find that implementing a bonus payment significantly reduces the underprovision observed under CAP ($p<0.0314$, FPPP, two-sided) and the overprovision observed under FFS ($p=0.0000$, FPPP, two-sided). Moreover, patients significantly benefit from introducing P4P incentives in CAP and in FFS as subjects in the role of physicians reduce their deviation from the patient optimal quantity to a similar extent ($p>0.5926$, FPPI, two-sided). The reduction of deviation is significantly affected by the severity of illness, though ($p<0.0002$, FPPP, two-sided). In particular, we observe the highest reduction for the most severe illness in CAP and the least severe illness in FFS, respectively.

Figure 7: Average Treatment Quantity per Severity without and with P4P



With P4P incentives, we still observe a significant underprovision under a CAP baseline payment (p=0.0000, FPPP, two-sided) and a significant overprovision under a FFS baseline payment (p=0.0000, FPPP, two-sided).[11] But, in contrast to the baseline incentives, the distances to the patient optimal level now only significantly differ across some severities in FFS P4P, i.e. between severities *x* and *z* (p=0.0019) and between severities *y* and *z* (p=0.0074 FPPP, two-sided; see Figure 8). The results are summarized in observation 2.

Figure 8: Distance between Patient Optimum and Chosen Quantity with P4P Incentives



Result 2: Patients benefit from introducing P4P incentives as these incentives significantly mitigate the underprovision (overprovision) observed with the baseline scheme CAP (FFS). Though, the deviations from the patient optimal treatment do not completely vanish with P4P.

3.3. Medical versus non-medical students

Our third research question analyzes whether differences in provision behavior exist between non-medical students – who are the typical subjects in laboratory experiments – and prospective physicians. In particular, we test the supposition that making a decision in a

[11] This result is supported by an OLS regression of the P4P incentive on the quantity of medical treatment, see Appendix A.1.

medical context might be induced by the different professional background and, accordingly leads to different decisions made by the two subject groups (see Ahlert et al., 2012).

Figure 9: Average Treatment Quantity per Severity for Medical and Non-medical Students

Figure 9 shows the average of treatment levels per severity for non-medical and medical students. There are no significant differences between the two subject groups in all payment schemes (p>0.1599, FPPI, two sided), except a weak one for severity *x* in FFS (p=0.0505, FPPT, two-sided). Only in the latter case we find that medical students are somewhat more patient-oriented than non-medical students. Moreover, all previous observations on the baseline payment schemes and the effect of P4P incentives also hold for prospective physicians.

Result 3: There are almost no significant differences in provision behavior between medical and non-medical students.

3.4. Robustness Check: Constant Profit Maximum

Introducing P4P incentives in our experiment is associated with an increase in physicians' maximum profit. We, therefore, test whether similar behavioral effects can be observed when keeping the profit maximum constant between the baseline and the P4P schemes.[12] In condition CAP12, we designed a CAP payment scheme in such a way that the physician profit

[12] As we find no differences between medical and other students, we pool the data from here on.

maximum is identical with and without P4P incentives, i.e. $\hat{\pi}$=12. We find no significant difference in average treatment quantities per severity provided between a CAP scheme with $\hat{\pi}$=12 (part 1 in condition CAP12) and a CAP scheme with $\hat{\pi}$=10 (part 1 in condition C1; p≥0.2397 FPPI, two-sided). Similarly, we find no significant difference between the CAP P4P schemes in part 2 of conditions CAP12 and C1, which have the same profit maximum of $\hat{\pi}$=12 (p>0.4405 FPPI, two-sided). Accordingly, patients benefit from introducing P4P incentives also when this payment reform is not associated with an increase of physicians' profit maximum.

Result 4: It appears that no increase in the maximum profit level is necessary to achieve the same patient health benefits.

This result raises the question on how patient benefits and the expenditure for physicians' payment are related to each other in the different conditions.

3.5. Ratio of Patient Benefit/Physician Remuneration

In this section, we analyze how patient benefits and the expenditure for physicians' remuneration are related in the different conditions. For this, we calculate the individual ratios *r* of the sum of benefits over the nine patients and the sum of respective remunerations and compare them between the different payment schemes.[13] As already noted, patient benefits significantly increase after introducing P4P incentives. But, expenses for physicians' remuneration also significantly increase with P4P incentives, particularly in conditions C1 and C2 (p<0.0140, FPPP, two-sided). Even in CAP12, where the profit maximum is the same in the baseline and in the P4P payment schemes, we find a significant increase in expenses (p=0.0002, FPPP, two-sided). We, thus, observe both, an increase in patient benefits and in physicians' remuneration. These results support recent field studies reporting that P4P schemes increase the quality of care, but come at increased costs (Mullen et al., 2010). Compared to these field studies, our controlled experimental environment allows clearly identifying the overall effect of P4P on the benefits per monetary unit spent (see Table 4).

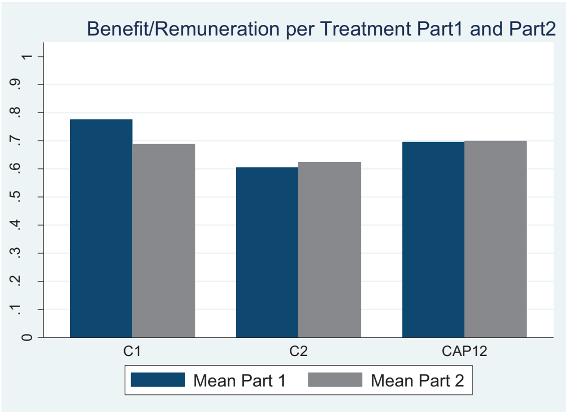
Table 4: Patient Benefit and Expenditure on Physician Remuneration per Condition

Condition	Part 1			Part 2		
	Patient Benefit	Expenditure on remuneration	Avg. *r*	Patient Benefit	Expenditure on remuneration	Avg. *r*
C1	7.7605	10.0000	0.7761	9.4074	13.7807	0.6876
C2	7.4656	14.2751	0.6056	9.3624	15.3016	0.6241
CAP12	8.3504	12.0000	0.6959	9.5983	13.8667	0.6991

[13] The minimum and maximum possible ratios for the four payment schemes given the nine patient types are: CAP (CAP P4P) r_{min}=0 (0), r_{max}=1.4 (1.13) and FFS (FFS P4P) r_{min}= 0 (0), r_{max}=5 (3.45). Note that the maximum levels for FFS (FFS P4P) are given for a quantity of one as for a quantity of zero the expenditure would be zero and thus the ratio undefined. However, we can exclude these cases as none of the subjects chose zero given FFS (FFS P4P).

Comparing the individual ratios *r* between the baseline and P4P schemes, we find that they do not differ significantly for C2 and CAP12 ($p>0.6136$, FPPT, two-sided), whereas they significantly decrease for C1 ($p=0.0040$, FPPP, two-sided). Thus, the patient benefits per monetary unit spent might even decrease with P4P incentives. Figure 10 illustrates the average benefit-expenditure ratios per condition before and after the P4P reform.

Figure 10: Average Benefit/Remuneration per Condition before and after the Reform



Result 5: The increase in patient benefits cannot overcompensate the additional expenditures associated with pay-for-performance incentives.

3.6. Individual Behavior

In this section, we focus on individual behavior and its changes between the two parts of the experiment. In particular, we identify four behavioral types: profit maximizing subjects (PMs) choosing their profit maximum quantity, benefit maximizing subjects (BMs) choosing the patient optimal quantity, trade-off types (TOs) choosing a quantity between the profit maximum and the patient optimal quantity, and others (Os) whose quantity choices cannot be explained by these types.[14] We define a subject to match one of these four behavioral types if the majority of his/her decisions in each part of the experiment is consistent with the classification criterion. If this is not the case we classify the subject as non-consistent (NCs).[15] The number of subjects per type in each condition and part of the experiment is given in Table 5. We pool our data over medical and non-medical students as the distribution of behavioral types across these subject groups is not significantly different (baseline: $p=0.834$, P4P incentives: $p=0.970$, two-sample Kolmogorov-Smirnov test; supporting our results in section 3.3). Similarly, we pool our data over (the symmetric) CAP and FFS schemes as we find also no significant effect across these schemes (baseline:

[14] Our design does not allow identifying social welfare optimizers, i.e. subjects who choose the quantity yielding the maximum sum of patient benefit and physician profit. The reason is that we cannot differentiate clearly between subjects who maximize social welfare and those who maximize patient benefits except for two of the nine patients for whom social welfare-maximizing quantities differ from patient optimal ones.

[15] This class comprises subjects that choose Pareto-inefficient quantities.

p=0.900, P4P incentives: p=0.970, two-sample Kolmogorov-Smirnov test; supporting our results in sections 3.1 and 3.2).

Table 5: Distribution of Patient Types per Part

	Baseline Payment	With P4P Incentives
PM	10	63
BM	26	33
TO	56	0
O	1	1
NC	7	3
Total	**100**	**100**

Our classification reveals that, without P4P incentives, 56 percent of subjects can be classified as TOs, 26 percent as BMs, and only 10 percent as PMs. After introducing P4P incentives, 63 percent of subjects can be classified as PMs. Also, the fraction of BMs is increased to 33 percent. Obviously, as the loss of profits associated with an optimal medical treatment decreases with P4P incentives, more subjects are willing to bear this lower loss to consistently treat their patients in an optimal way. Since the distance between the profit maximum quantity and the benefit maximum quantity is only one unit in part 2, it is not possible to identify TOs in this part. While, in part 1, there are 7 percent of subjects whose behavior is not consistent with one of the four types, in part 2 the percentage of NCs is decreased to 3 percent. Only 1 subject (1 percent of overall decisions) does not follow the behavioral pattern of PM, BM, or TO.

The analysis of individual changes of behavior between the two parts reveals a more detailed picture of how P4P incentives work (see Table 6). We find that only 29 percent of subjects do not change their type. These are particularly PMs (90 percent of them do not change their type) and BMs (77 percent of them do not change their type). For all other subjects (71 percent) the type classification changes with P4P incentives. The majority of them are TOs (56 out of 71) who mainly switch to either PM (45) or, to a minor degree, BM (8). The six BMs who change their behavior, switch to PM. The latter finding indicates that the monetary P4P incentives might crowd out the subjects' intrinsic motivation to maximize patient benefits, at least to some degree.

Table 6: Individual Type Changes from Part 1 to Part 2 per Patient

Type Changes	Number of Subjects
PM-PM	9
PM-O	1
BM-BM	20
BM-PM	6
TO-PM	45
TO-BM	8
TO-NC	3
O-BM	1
NC-PM	4
NC-BM	3
Total	**100**

Another way to classify subjects is to use a measure averaging the treatment behavior over all patients. We base our classifications on the individual absolute distance between the chosen quantity and the patient optimal quantity averaged over all nine patients per part (d_i).[16] The classifications are derived as follows.

First, we consider the classification of BMs. The patient optimal quantity averaged over all nine patients is $q_{kl}^* = 5$ irrespective of the payment scheme. Accordingly, patient optimal treatment implies that the chosen quantities averaged over all nine patients do not deviate from $q_{kl}^* = 5$, i.e. the patient optimal distance over all nine patients is $d_{kl}^* = 0$. To be classified as a BM we allow for an individual deviation of less than 0.5 from the patient optimal distance, i.e, $d_i < 0.5$.

Second, we derive the classification of PMs. With baseline incentives in part 1, the profit maximizing quantity for all nine patients is $\hat{q}_{kl} = 0$ for CAP and $\hat{q}_{kl} = 10$ for FFS. Thus, the profit maximizing distance to the patient optimal quantity ($q_{kl}^* = 5$) is $\hat{d}_{kl} = 5$ under both baseline schemes. With P4P incentives in part 2, the profit maximizing quantity for all nine patients is $\hat{q}_{kl} = 4$ for CAP P4P and $\hat{q}_{kl} = 6$ for FFS P4P. Hence, the profit maximizing distance to the patient optimal quantity ($q_{kl}^* = 5$) is $\hat{d}_{kl} = 1$ for both P4P schemes, We keep the classification of PMs comparable between the two schemes insofar as the thresholds, i.e. the allowed deviations from the optimal distances, are the same in both schemes. With baseline incentives in part 1, subjects are classified as PMs, if $d_i > 4.5$. With P4P incentives in part 2, subjects are classified as PMs, if $d_i > 0.5$.

Third, we consider TOs who trade-off patient benefits with own profits. Their individual distances are in between the ones of BMs and PMs. Hence, with baseline incentives in part 1 a subject is classified as TO, if $0.5 \leq d_i \leq 4.5$, and with P4P incentives in part 2 a subject is classified as TO, if $d_i = 0.5$.[17]

[16] Due to the symmetric design of the benefit and payment functions in part 1 and part 2, respectively, this measure does not need to be adjusted between CAP (CAP P4P) and FFS (FFS P4P).

[17] Note that given these classifications social welfare optimizers are identical to TOs.

Table 7: Distribution of Patient Types per Part

	Baseline Payment	With P4P Incentives
PM	10	67
BM	16	33
TO	74	-
Total	**100**	**100**

Given this classification, we find that, in part 1, 74 percent of subjects are TOs, 16 percent are BMs, and only 10 percent are PMs (see Table 7). This distribution of types is very similar to the one based on individual decisions. The percentages change when a P4P bonus is introduced in part 2 of the experiment. With P4P incentives, 67 percent of subjects can be classified as PMs and 33 percent as BMs. The shift toward profit maximizing behavior, again, underlines the effectiveness of financial incentives. At the same time we see that the lower loss of profits associated with an optimal medical treatment of patients induces more subjects to behave in this way.

Again, the comparison of the individual classification between the two parts reveals a more detailed picture of the functioning of P4P incentives (see Table 8). We find that, with this second classification, a similar percentage of subjects (25 percent) who do not change their type. Again, these are particularly PMs (100 percent of them do not change their type) and BMs (94 percent of them do not change their type). All other subjects (75 percent) do change their type classification with P4P incentives. The majority of them are TOs (74 out of 75) who mainly switch to either PM (56) or, to a minor degree, BM (18). The one BM who changes her behavior, switches to PM. That is, the insights generated with our first classification also hold for the second one. Note that with this second classification we find almost no hint for a crowding out of the intrinsic motivation to treat the patients in an optimal way.

Table 8: Switching behavior of Types between Part 1 and Part 2 over all Patients

Type Changes	Number of Subjects
BM-BM	15
BM-PM	1
PM-PM	10
TO-BM	18
TO-PM	56
Total	**100**

Result 6: Profit maximizing subjects and benefit maximizing subjects reveal a rather stable behavior between the two parts of the experiment. Particularly those, who trade off the own profit with the patient benefit, change their behavior with P4P incentives. The majority of them switches to profit maximization (which comes along with an increase in patient benefits).

4. Conclusion

The aim of this study is to analyze P4P incentives within a laboratory experiment. In contrast to a large part of the existing research, the laboratory environment allows isolating the effects of financial P4P incentives from other influencing factors and, thus, drawing conclusions on how P4P schemes work. In particular, the experimental setting allows introducing a clear-cut performance measure directly linked to the patient's benefit which is different from many field studies. Our data reveal that physician provision behavior under P4P schemes differs significantly from behavior under more traditional fee-for-service and capitation schemes. In particular, being paid a bonus for good performance, physicians provide patients with medical services much closer to their optimal treatment levels, hence improving the quality of care. The observed effects hold for both non-medical students (the pre-dominant population analyzed in laboratory experiments) and prospective physicians.

Our results lead us to conclude that in case quality metrics are well designed in order to prevent other problems observed in the field such as substitution from non-incentivized to incentivized aspects of medical care (Eggleston, 2005, Kaarboe and Siciliani, 2011), patient exclusion (Gravelle et al., 2010), or reliability of evidence-based outcome measures (Maynard, 2012), P4P incentives may actually lead to better patient outcomes.

However, comparing the benefit/expenditure ratio of payment schemes without and with P4P incentives, we observe in almost all treatments that this ratio is not significantly improved when P4P is introduced. In one case it even slightly decreases. From a welfare economic perspective, the increase in patient health benefits, thus, cannot overcompensate the additional expenditure due to P4P incentives in our parameter setting. This result is in line with findings of previous field studies (see e.g. Mullen et al. 2010). Hence, in case policy makers aim at improving the quality of care in terms of better treatment of patients, P4P incentives can be a successful policy means. However, if policy makers also want to improve the patient benefit per monetary unit spent, P4P incentives alone may not be sufficient. This insight holds even if the P4P scheme is designed in a way that the profit maximum is held constant compared to the baseline payment scheme. This may be an important indicator for policy makers in terms of money needed to provide P4P incentives.

Finally, our analysis of individual switching behavior supports the effectiveness of financial P4P incentives. While the majority of physicians trades off patient benefits against their own payoff under the basic payment schemes in part 1 of our experiment, most of them become profit maximizers in part 2 which goes along with the targeted increase in patient benefits. That is, aligning financial incentives for physicians with patient health benefits is a successful means to bring medical treatment closer to the patient optimum. Though, in some individual cases financial incentives might also crowd out the intrinsic motivation to treat patients well.

Literature

Ahlert, M., Felder, S., and B. Vogt (2012): Which Patients Do I Treat?: An Experimental Study with Economists and Physicians, *Health Economics Review*, 2(1), 1–11.

Baltussen, G., Post, G. T., Van Den Assem, M. J., and P.P. Wakker, (2012): Random incentive systems in a dynamic choice experiment. *Experimental Economics*, *15*(3), 418-443.

Borowitz, M., Scheffler, R.M. and B.M., Fulton (2010): Improving value for money in health by paying for performance. In: OECD Health Policy Studies, Value for Money in Health Spending; Paris: Organisation for Economic Co-operation and Development.

Brosig-Koch, J., H. Hennig-Schmidt, N. Kairies, D. Wiesen (2013): How to improve patient care? An analysis of capitation, fee-for-service, and mixed incentive schemes for physicians, *Ruhr Economic Papers No*. 412 .

Buetow, S. (2008): Pay-for-Performance in New Zealand Primary Health Care, *Journal of Health Organisation and Management*, 22, 36-47.

Campbell, S.M, D. Reeves, E. Kontopantelis, B. Sibbald, and M. Roland (2009): Effects of Pay for Performance on the Quality of Primary Care in England, *New England Journal of Medicine*, 361, 368-78.

Choné, P. and C.-t.A. Ma (2010): Asymmetric Information from Physician Agency: Optimal Payment and Health Care Quantity, Working Paper, University of Boston.

Cubitt, R. P., Starmer, C., and R. Sugden (1998): On the validity of the random lottery incentive system, *Experimental Economics*, 1(2), 115-131.

Donabedian, A. (2005): Evaluating the Quality of Medical Care, *The Milbank Quarterly*, 83 (4), 691-729.

Doran, T., C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland (2006): Pay-for-Performance Programs in Family Practices in the United Kingdom, *New England Journal of Medicine*, 355, 375-384.

Duckett, S., Daniels, S., Kamp, M., Stockwell, A., and G. Walker (2008): Pay for Performance in Australia: Queensland's new Clinical Practice Improvement Payment, *Journal of Health Services Research and Policy*. 13, 174-177.

Eckel, C. and P. Grossman, (1996): Altruism in anonymous dictator games, *Games and Economic Behavior*, 16, 181.

Eggleston, K. (2005): Multitasking and Mixed Payment Systems for Provider Payment, *Journal of Health Economics* 24, 211-223.

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

Fischbacher, U. (2007): Z-tree: Zurich Toolbox for Readymade Economic Experiments, *Experimental Economics*, 10, 171-178.

Frey, B. S. and F. Oberholzer-Gee (1997): The cost of price incentives: An empirical analysis of motivation crowding-out, *The American economic review*, *87*(4), 746-755.

Frey, B. S. and R. Jegen (2001): Motivation crowding theory, *Journal of economic surveys*, 15(5), 589-611.

Gené-Badia, J., Escaramis-Babiano, G., Sans-Corrales, M., Sampietro-Colom, L., Aguado-Menguy, F., Cabezas-Pena, C., and Gallo-Puelles, P. (2007): Impact of economic incentives on quality of professional life and on end-user satisfaction in primary care, *Health Policy*, 80, 2–10.

Gneezy, U. and A. Rustichini (2000): Pay Enough or Don’t Pay at All, *Quarterly Journal of Economics*, 115, 791–810.

Gravelle, H., Sutton, M., and A. Ma (2010): Doctor behaviour under a pay for performance contract: treating, cheating and case finding?, *The Economic Journal* 120, 129–156.

Greiner, B. (2004): An Online Recruitment System for Economic Experiments, in V. M. Kurt Kremer (ed.), Forschung und Wissenschaftliches Rechnen 2003. GWDG Bericht 63, Gesellschaft für Wissenschaftliche Datenverarbeitung, Göttingen, 79-93.

Hennig-Schmidt, H., Selten, R., and D. Wiesen (2011): How Payment Systems Affect Physicians' Provision Behavior? An Experimental Investigation, *Journal of Health Economics*, 30(4), 637-646.

Kaarboe, O. and L. Siciliani, (2011): Multi-tasking, quality and pay for performance, *Health Economics*, 20(2), 225-238.

Kim, B.Y. (2010): Quality Assessment and Pay for Performance in Korea, HIRA, Korean Health Insurance Review Agency.

Laury, S.K. (2006): Pay One or Pay All: Random Selection of One Choice for Payment, Experimental Economics Center Working Paper Series, No 2006-24, Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University.

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

Maynard, A. (2012): The powers and pitfalls of payment for performance, *Health Economics*, 21 (1), 3-12.

McClellan, M. (2011): Reforming payments to healthcare providers: The key to slowing healthcare cost growth while improving quality?, *Journal of Economic Perspectives*, 25, 69-92.

Mellström, C. and M. Johannesson (2008). Crowding out in blood donation: was Titmuss right?, *Journal of the European Economic Association*, *6*(4), 845-863.

Mousquès, J., Afritte, A., and Y. Bourgueil (2012): Efficiency of Multidisciplinary Group Practices in Primary Care – An assessment of a French Experiment of Pay for Performance for Integrated Primary Care Organization, mimeo, ECHE.

Mullen, K., Frank, R., and M. Rosenthal (2010): Can You Get What You Pay For? Pay-For-Performance and the Quality of Healthcare Providers, *The RAND Journal of Economics*, 41(1), 64-91.

Perkins, R. and M. Seddon, on behalf of EPIQ, (2006): Quality improvement in New Zealand healthcare. Part 5: measurement for monitoring and controlling performance–the quest for external accountability, New Zealand Medical Journal, 119.

Rosenthal, M.B., Fernandopulle, R., Song, H.R., and B. Landon (2004): Paying for Quality: Providers' Incentives for Quality Improvement, *Health Affairs*, 23 (2), 127-41.

Rosenthal, M.B., (2008): Beyond Pay for performance – Emerging Models of Provider Payment Reform, *New England Journal of Medicine*, 359, 1197-1200.

Scott, I. (2008): Pay for Performance Programs in Australia: A Need for Guiding Principles, *Australian Health Review*, 32, 740–9.

Scott, A., Schurer, S., Jensen P.H, and P. Sivey (2009): The Effects of an Incentive Program of Quality of Care in Diabetes Management, Health Economics, 9, 1091–108.

Starmer, C. and R. Sugden (1991): Does the random-lottery incentive system elicit true preferences? An experimental investigation, *The American Economic Review*, 81(4), 971-978.

APPENDICES

A.1. OLS Regression: P4P incentive on the quantity of medical treatment clustered by subject

VARIABLES	C1	C2
P4P incentive	1.546***	-1.789***
	(0.366)	(0.283)
Constant	1.285	9.294***
	(0.798)	(0.620)
Observations	414	360
R-squared	0.124	0.216
N_clust	23	20

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

A.2 Instructions and Control Questions

Welcome to the Experiment!

Preface

You are participating in an economic experiment on decision behavior. You and the other participants will be asked to make decisions for which you can earn money. Your payoff depends on the decisions you make. At the end of the experiment, your payoff will be converted to Euro and paid to you in cash. During the experiment, all amounts are presented in the experimental currency Taler. 10 Taler equals 8 Euro.

The experiment will take about 90 minutes and consists of two parts. You will receive detailed instructions before each part. Note that none of your decisions in either part have any influence on the other part of the experiment.

Part One

Please read the following instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you.

Part one of the experiment consists of 9 rounds of decision situations.

Decision Situations

In each round you take on the role of a physician and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness.

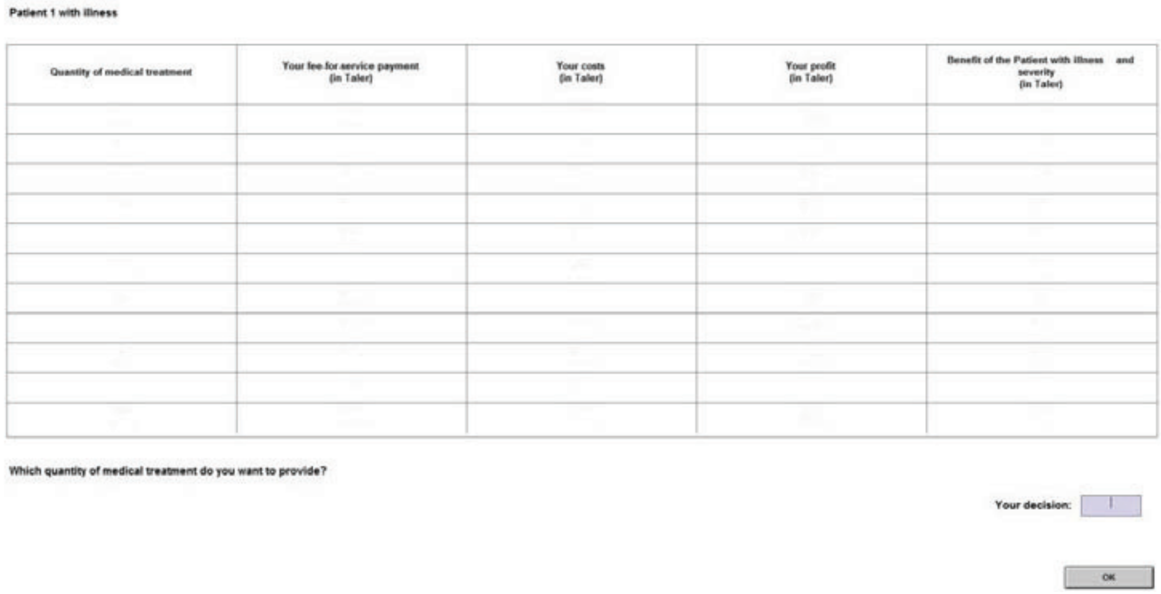
Every patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Profit

In each round you receive a fee-for-service (capitation) remuneration for treating the patient. Your remuneration increases with the amount of medical treatment (is irrespective of the amount of medical treatment) you provide. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the fee-for-service (capitation) remuneration.

Every quantity of medical service yields a particular benefit for the patient – contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit but also the patient’s benefit.

In each round you will receive detailed information on your screen (see below) for the respective patient the illness, your amount of fee-for-service (capitation) remuneration - for each possible amount of medical treatment - your costs, profit as well as the benefit for the patient with the corresponding illness and severity.



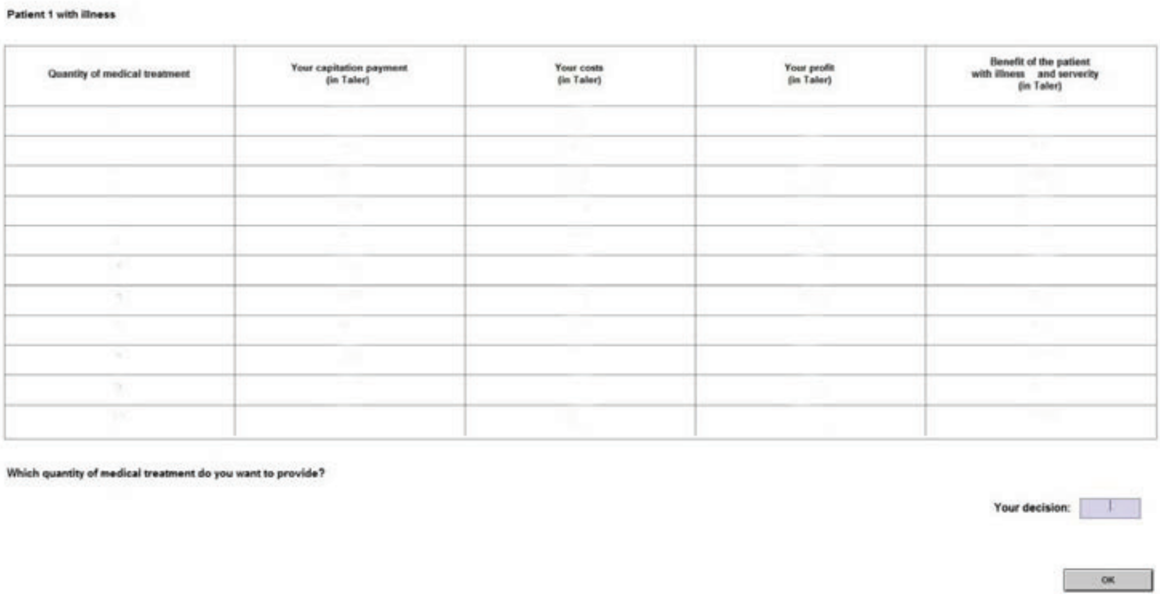
Patient 1 with illness

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the Patient with illness and severity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

OK



Patient 1 with illness

Quantity of medical treatment	Your capitation payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and serverity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

OK

Payment

At the end of the experiment one of the 9 rounds of part one will be chosen at random. Your profit in this round will be paid to you in cash.

For this part of the experiment, no patients are physically present in the laboratory. Yet, the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, an organization which funds the treatment of patients with eye cataract.

The transfer of money to the Christoffel Blindenmission Deutschland e.V. will be carried out after the experiment by the experimenter and one participant. The participant completes a money transfer form, filling in the total patient benefit (in Euro) resulting from the decisions made by all participants in the randomly chosen situation. This form prompts the payment of the designated amount to the Christoffel Blindenmission Deutschland e.V. by the University of Duisburg-Essen's finance department. The form is then sealed in a postpaid envelope and posted in the nearest mailbox by the participant and the experimenter.

After the entire experiment is completed, one participant is chosen at random to oversee the money transfer to the Christoffel Blindenmission Deutschland e.V. The participant receives an additional compensation of 5 Euro for this task. The participant certifies that the process has been completed as described here by signing a statement which can be inspected by all participants at the office of the Chair of Quantitative Economic Policy. A receipt of the bank transfer to the Christoffel Blindenmission Deutschland e.V. may also be viewed here.

Comprehension Questions

Prior to the decision rounds we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part one of the experiment will begin once all participants have answered the comprehension questions correctly.

Part II

Please read the following instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you.

Part two of the experiment also consists of 9 rounds of decision situations.

Decision Situations

As in part one of the experiment, you take on the role of a physician in each round and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness.

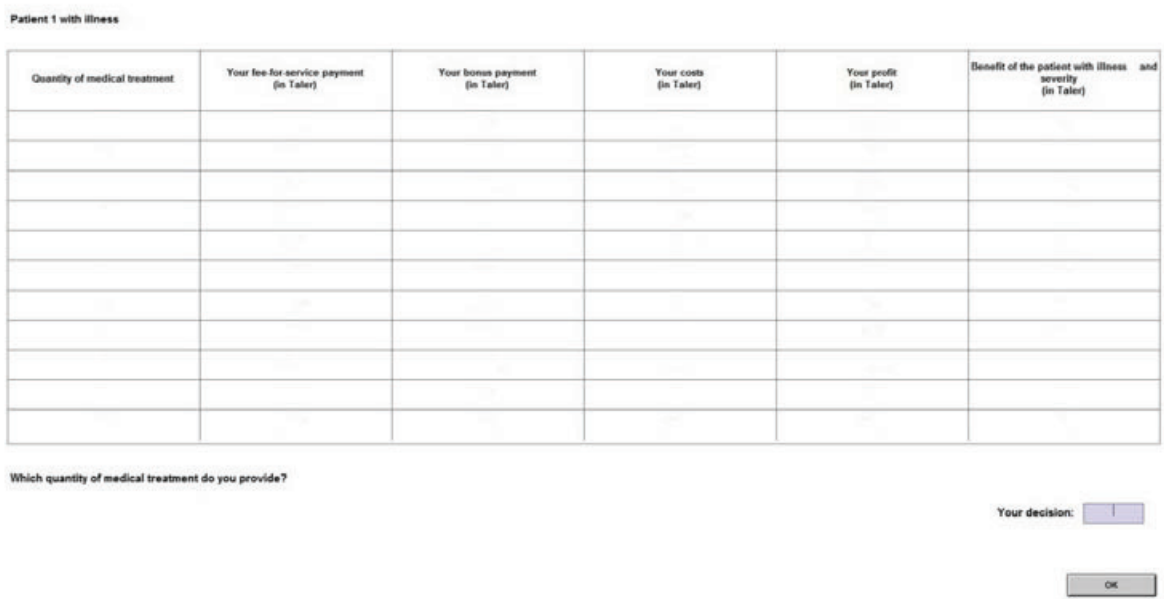
Every patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Profit

In each round you are remunerated for treating the patient. In each round you receive a fee-for-service (capitation) remuneration for treating the patient. Your remuneration increases with the amount of medical treatment (is irrespective of the amount of medical treatment) you provide. In addition to this, in each round you receive a bonus payment, in case the quantity of medical services you provide is equal to the one that results in the highest benefit for the patient, or deviates by one quantity from the latter. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the sum of your fee-for-service (capitation) remuneration and bonus payment.

As in part one, every quantity of medical service yields a particular benefit for the patient – contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit but also the patient's benefit.

In each round you will receive detailed information on your screen (see below) for the respective patient the illness, your amount of fee-for-service (capitation) remuneration - for each possible amount of medical treatment - the amount of your bonus payment, your costs, profit as well as the benefit for the patient with the corresponding illness and severity.



Patient 1 with illness

Quantity of medical treatment	Your capitation payment (in Taler)	Your bonus payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the patient with illness and severity (in Taler)

Which quantity of medical treatment do you want to provide?

Your decision:

OK

Payment

At the end of the experiment one of the 9 rounds of part two will be chosen at random. Your profit in this round will be paid to you in cash, in addition to your payment from the round chosen for part one of the experiment.

After the experiment is over, please remain seated until the experimenter asks you to step forward. You will receive your payment at the front of the laboratory before exiting the room.

As in part one, no patients are physically present in the laboratory for part two of the experiment. Yet, the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, an organization which funds the treatment of patients with eye cataract.

The process for the transfer of money to the Christoffel Blindenmission Deutschland e.V. as described for part one of the experiment will be carried out by the experimenter and one participant.

Comprehension Questions

Prior to the decision rounds we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part two of the experiment will begin once all participants have answered the comprehension questions correctly.

Finally, we kindly ask you to not talk to anyone about the content of this session in order to prevent influencing other participants after you. Thank you for your collaboration!

Comprehension Questions Part I: CAP (FFS)

Questions Tables 1-4:

1-4 a) What is the capitation (fee-for-service)?

1-4 b) What are the costs?

1-4 c) What is the profit?

1-4 d) What is the patient benefit?

Quantity of medical treatment	Capitation (Fee-for-service) (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	20.00 (0.00)	0.00	20.00 (0.00)	15.00
1	20.00 (4.00)	0.20	19.80 (3.80)	16.00
2	20.00 (8.00)	0.80	19.20 (7.20)	17.00
3	20.00 (12.00)	1.80	18.20 (10.20)	18.00
4	20.00 (16.00)	3.20	16.80 (12.80)	19.00
5	20.00 (20.00)	5.00	15.00 (15.00)	20.00
6	20.00 (24.00)	7.20	12.80 (16.80)	19.00
7	20.00 (28.00)	9.80	10.20 (18.20)	18.00
8	20.00 (32.00)	12.80	7.20 (19.20)	17.00
9	20.00 (36.00)	16.20	3.80 (19.80)	16.00
10	20.00 (40.00)	20.00	0.00 (20.00)	15.00

1. Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.
2. Assume that a physician wants to provide 9 quantities of medical treatment for the patient depicted above.

Quantity of medical treatment	Capitation (Fee-for-service) (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness G and severity z (in Taler)
0	20.00 (0.00)	0.00	20.00 (0.00)	10.00
1	20.00 (4.00)	0.20	19.80 (3.80)	12.00
2	20.00 (8.00)	0.80	19.20 (7.20)	14.00
3	20.00 (12.00)	1.80	18.20 (10.20)	16.00
4	20.00 (16.00)	3.20	16.80 (12.80)	18.00
5	20.00 (20.00)	5.00	15.00 (15.00)	20.00
6	20.00 (24.00)	7.20	12.80 (16.80)	22.00
7	20.00 (28.00)	9.80	10.20 (18.20)	24.00
8	20.00 (32.00)	12.80	7.20 (19.20)	22.00
9	20.00 (36.00)	16.20	3.80 (19.80)	20.00
10	20.00 (40.00)	20.00	0.00 (20.00)	18.00

3. Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.
4. Assume that a physician wants to provide 9 quantities of medical treatment for the patient depicted above.

Comprehension Questions Part II: CAP(FFS) P4P

1-4 a) What is the capitation (fee-for-service)?

1-4 a) What is the bonus payment?

1-4 b) What are the costs?

1-4 c) What is the profit?

1-4 d) What is the patient benefit?

Quantity of medical treatment	Capitation (Fee-for-service (in Taler)	Bonus payment (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	20.00 (0.00)	0.00 (0.00)	0.00	20.00 (0.00)	15.00
1	20.00 (4.00)	0.00 (0.00)	0.20	19.80 (3.80)	16.00
2	20.00 (8.00)	0.00 (0.00)	0.80	19.20 (7.20)	17.00
3	20.00 (12.00)	0.00 (0.00)	1.80	18.20 (10.20)	18.00
4	20.00 (16.00)	7.20 (7.20)	3.20	24.00 (20.00)	19.00
5	20.00 (20.00)	7.20 (7.20)	5.00	22.20 (22.20)	20.00
6	20.00 (24.00)	7.20 (7.20)	7.20	20.00 (24.00)	19.00
7	20.00 (28.00)	0.00 (0.00)	9.80	10.20 (18.20)	18.00
8	20.00 (32.00)	0.00 (0.00)	12.80	7.20 (19.20)	17.00
9	20.00 (36.00)	0.00 (0.00)	16.20	3.80 (19.80)	16.00
10	20.00 (40.00)	0.00 (0.00)	20.00	0.00 (20.00)	15.00

1. Assume that a physician wants to provide 1 quantities of medical treatment for the patient depicted above.
2. Assume that a physician wants to provide 8 quantities of medical treatment for the patient depicted above.

Quantity of medical treatment	Capitation (Fee-for-service (in Taler)	Bonus payment (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness G and severity z (in Taler)
0	20.00 (0.00)	0.00 (0.00)	0.00	20.00 (0.00)	10.00
1	20.00 (4.00)	0.00 (0.00)	0.20	19.80 (3.80)	12.00
2	20.00 (8.00)	0.00 (0.00)	0.80	19.20 (7.20)	14.00
3	20.00 (12.00)	0.00 (0.00)	1.80	18.20 (10.20)	16.00
4	20.00 (16.00)	0.00 (0.00)	3.20	16.80 (12.80)	18.00
5	20.00 (20.00)	0.00 (0.00)	5.00	15.00 (15.00)	20.00
6	20.00 (24.00)	11.20 (4.80)	7.20	24.00 (21.60)	22.00
7	20.00 (28.00)	11.20 (4.80)	9.80	21.40 (23.00)	24.00
8	20.00 (32.00)	11.20 (4.80)	12.80	18.40 (24.00)	22.00
9	20.00 (36.00)	0.00 (0.00)	16.20	3.80 (19.80)	20.00
10	20.00 (40.00)	0.00 (0.00)	20.00	0.00 (20.00)	18.00

3. Assume that a physician wants to provide 1 quantities of medical treatment for the patient depicted above.
4. Assume that a physician wants to provide 8 quantities of medical treatment for the patient depicted above.

A.3 Parameter Tables

Treatment	Variable	Quantity (q) 0	1	2	3	4	5	6	7	8	9	10
C1	R_{kl}^{I}	10	10	10	10	10	10	10	10	10	10	10
	R_{kl}^{II}	10	10	10	10	10	10	10	10	10	10	10
	$P_{x}^{II\ CAP}$	0	0	2.4	2.4	2.4	0	0	0	0	0	0
	$P_{y}^{II\ CAP}$	0	0	0	0	3.6	3.6	3.6	0	0	0	0
	$P_{z}^{II\ CAP}$	0	0	0	0	0	0	5.6	5.6	5.6	0	0
C2	R_{kl}^{I}	0	2	4	6	8	10	12	14	16	18	20
	$R_{kl}^{II\ FFS}$	0	2	4	6	8	10	12	14	16	18	20
	$P_{x}^{II\ FFS}$	0	0	5.6	5.6	5.6	0	0	0	0	0	0
	$P_{y}^{II\ FFS}$	0	0	0	0	3.6	3.6	3.6	0	0	0	0
	$P_{z}^{II\ FFS}$	0	0	0	0	0	0	2.4	2.4	2.4	0	0
CAP12	R_{kl}^{I}	12	12	12	12	12	12	12	12	12	12	12
	R_{kl}^{II}	10	10	10	10	10	10	10	10	10	10	10
	$P_{x}^{II\ CAP}$	0	0	2.4	2.4	2.4	0	0	0	0	0	0
	$P_{y}^{II\ CAP}$	0	0	0	0	3.6	3.6	3.6	0	0	0	0
	$P_{z}^{II\ CAP}$	0	0	0	0	0	0	5.6	5.6	5.6	0	0
all	c_{kl}	0	0.1	0.4	0.9	1.6	2.5	3.6	4.9	6.4	8.1	10
C1	π_{kl}^{I}	10	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6	1.9	0
	π_{x}^{II}	10	9.9	12	11.5	10.8	7.5	6.4	5.1	3.6	1.9	0
	π_{x}^{II}	10	9.9	9.6	9.1	12	11.1	10	5.1	3.6	1.9	0
	π_{x}^{II}	10	9.9	9.6	9.1	8.4	7.5	12	10.7	9.2	1.9	0
C2	π_{kl}^{I}	0	1.9	3.6	5.1	6.4	7.5	8.4	9.1	9.6	9.9	10
	π_{x}^{II}	0	1.9	9.2	10.7	12	7.5	8.4	9.1	9.6	9.9	10
	π_{x}^{II}	0	1.9	3.6	5.1	10	11.1	12	9.1	9.6	9.9	10
	π_{x}^{II}	0	1.9	3.6	5.1	6.4	7.5	10.8	11.5	12	9.9	10
CAP12	π_{kl}^{I}	12	11.9	11.6	11.1	10.4	9.5	8.4	7.1	5.6	3.9	2
	π_{x}^{II}	10	9.9	12	11.5	10.8	7.5	6.4	5.1	3.6	1.9	0
	π_{x}^{II}	10	9.9	9.6	9.1	12	11.1	10	5.1	3.6	1.9	0
	π_{x}^{II}	10	9.9	9.6	9.1	8.4	7.5	12	10.7	9.2	1.9	0
all	B_{Ax}	4	5	6	7	6	5	4	3	2	1	0
	B_{Ay}	2	3	4	5	6	7	6	5	4	3	2
	B_{Az}	0	1	2	3	4	5	6	7	6	5	4
	B_{Bx}	7	8	9	10	9	8	7	6	5	4	3
	B_{By}	5	6	7	8	9	10	9	8	7	6	5
	B_{Bz}	3	4	5	6	7	8	9	10	9	8	7
	B_{Cx}	8	10	12	14	12	10	8	6	4	2	0
	B_{Cy}	4	6	8	10	12	14	12	10	8	6	4
	B_{Cz}	0	2	4	6	8	10	12	14	12	10	8

Chapter 4

How do Non-Monetary Performance Incentives for Physicians Affect the Quality of Medical Care? A Laboratory Experiment

Reference

Kairies, N. and M. Krieger. 2013. "How do Non-Monetary Performance Incentives for Physicians Affect the Quality of Medical Care? A Laboratory Experiment." *Ruhr Economic Papers* No. 414.

RUHR

ECONOMIC PAPERS

Nadja Kairies

Miriam Krieger

How do Non-Monetary Performance Incentives for Physicians Affect the Quality of Medical Care?

A Laboratory Experiment

UNIVERSITÄT DUISBURG ESSEN

#414

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #414

Responsible Editor: Volker Clausen

ISSN 1864-4872 (online) – ISBN 978-3-86788-469-3

Ruhr Economic Papers #414

Nadja Kairies and Miriam Krieger

How do Non-Monetary Performance Incentives for Physicians Affect the Quality of Medical Care?

A Laboratory Experiment

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen National-bibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

http://dx.doi.org/10.4419/86788469
ISSN 1864-4872 (online)
ISBN 978-3-86788-469-3

Nadja Kairies and Miriam Krieger[1]

How do Non-Monetary Performance Incentives for Physicians Affect the Quality of Medical Care? – A Laboratory Experiment

Abstract

In recent years, several countries have introduced non-monetary performance incentives for health care providers to improve the quality of medical care. Evidence on the effect of non-monetary feedback incentives, predominantly in the form of public quality reporting, on the quality of medical care is, however, ambiguous. This is often because empirical research to date has not succeeded in distinguishing between the effects of monetary and non-monetary incentives, which are usually implemented simultaneously. We use a controlled laboratory experiment to isolate the impact of non-monetary performance incentives: subjects take on the role of physicians and make treatment decisions for patients, receiving feedback on the quality of their treatment. The subjects' decisions result in payments to real patients. By giving either private or public feedback we are able to disentangle the motivational effects of self-esteem and social reputation. Our results reveal that public feedback incentives have a significant and positive effect on the quality of care that is provided. Private feedback, on the other hand, has no impact on treatment quality. These results hold for medical students and for other students.

JEL Classification: I11, C91, L15, I18

Keywords: Laboratory experiment; quality reporting; feedback; treatment quality; performance incentives

April 2013

1 Both CINCH and University of Duisburg-Essen. – We thank Jeannette Brosig-Koch, Stefan Felder, and the participants of the 2013 Conference of the German Association for Health Economics dggö for their valuable comments on this paper. Financial support from CINCH and from the German Research Foundation (DFG) is gratefully acknowledged. *– All correspondence to Nadja Kairies, University of Duisburg-Essen, CINCH and Faculty of Economics and Business Administration, Universitätsstr. 12, 45117 Essen, Germany, E-Mail: Nadja.Kairies@ibes.uni-due.de.*

1. Introduction

Recent healthcare reforms in various countries have specifically aimed at improving the quality of medical care while simultaneously controlling costs (McCellan, 2011). In this context non-monetary performance incentives, predominantly in the form of public quality reporting, have gained increasing popularity among policy makers as a means to achieving these two seemingly contrary goals (Dranove and Jin, 2010).

Evidence on public quality reporting in medical care shows that while it leads physicians to change their provision behavior (Kolstad, 2013), it is not clear that this actually improves the quality of care (Marshall, 2000). On the one hand, there is some indication that public quality reporting can contribute to decreased mortality rates; see Hannan et al. (1994) or Rosenthal et al. (1997). On the other hand, public reporting can also lead to unintended problems, such as a shift in effort towards those aspects of medical care that are reported on and away from unreported aspects (Werner et al., 2009), or the selection of patients towards those whose treatment improves the reported outcomes (Dranove et al., 2003, Cutler et al., 2004, Werner and Asch, 2005). Another issue is that regional characteristics influence the effects of public quality reporting on the quality of medical care: public quality reports have a larger impact, for instance, the more competitive the health care market is (Grabowski and Town, 2011). A further issue with previous empirical studies on the effects of non-monetary performance incentives in health care is that they are often difficult to disentangle from those of monetary incentives, especially as these two mechanisms are typically implemented together, e.g. in the US Premier Hospital Quality Incentive Demonstration within Medicare and in the UK Quality and Outcomes Framework. Simultaneous implementation of multiple new incentives makes it very difficult to establish what the individual impact of each of these changes to the system is, and whether they are in fact substitutes or complements (Maynard, 2012). Hence, Cutler et al. (2004) point out that more research on such incentives is essential to understanding the underlying mechanisms that drive changes in physician provision behavior.

From a theoretical point of view, it is important to differentiate between the modes of performance incentives, i.e. whether feedback is given privately or in public. Private feedback is a competitive incentive which addresses an individual’s self-esteem. Bénabou and Tirole (2002) state that the mere possibility of receiving positive feedback can motivate an individual to increase his performance. Making someone’s relative performance known to others, however, adds a reputational or image aspect to the incentive and speaks to the individual’s desire to gain social status and avoid social disapproval (see Bénabou and Tirole, 2006). On the other hand, monitoring performance and giving feedback also implies control, which can potentially crowd out pro-social behavior (see Ellingsen and Johannesson, 2008).

So far there is barely any empirical research specific to the health care sector which distinguishes the effect of (private) performance feedback based on self-esteem from that (public) based on social reputation. Hibbard et al. (2003) and Hibbard et al. (2005) report on an experimental field study in which they examine the effects of private as well as public feedback on the quality of care in hospitals. Their design includes two intervention groups, one of which receives both private and public feedback and one only private feedback, and a control group which is given no feedback at all. They find that hospitals which receive public feedback are significantly more involved in quality improvement efforts than hospitals with only private or with no feedback (Hibbard et al., 2003). Hospitals in both treatment groups increase their quality compared to the control group hospitals with no feedback, although the differences in average performance changes between the two treatment groups were not statistically significant (Hibbard et al., 2005). However, the results of these studies are subject to some methodological limitations, such as non-random assignment of hospitals to the groups and reliance on self-reported performance measures.

In non-medical settings, evidence for the positive impact of private feedback on performance is provided by several laboratory experiments, for instance Charness et al. (2011) and Kuhnen and Tymula (2012) for output in real-effort tasks. The positive impact of rank information on performance has also been documented in various field studies, including Mas and Moretti (2009) among factory shift workers, Azmat and Iriberri (2010) for high school students, and Blanes i Vidal and Nossol (2011) for white-collar workers. However, studies by Hannan et al. (2008), Eriksson et al. (2009), and Barankay (2011a and 2011b) all report results from laboratory or field experiments which suggest that feedback affects performance either negatively (at least for some individuals or under some conditions) or not at all. There is also evidence from laboratory experiments for a positive impact of public feedback on performance in non-medical settings, such as contributions to a public good (Rege and Telle, 2004) or donations to charity (Ariely et al., 2009). This effect has also been found in the field: In a study of Vietnamese language students, Tran and Zeckhauser (2012) report that both private and public feedback significantly raise test performance as compared to giving no feedback at all. In fact, students who were given public feedback outperformed those who received private feedback, though the difference is only marginally significant.

The lack of evidence for the health care market and the ambiguous results of performance feedback in other domains obscure the picture of how feedback incentives might work in a medical setting. The relevant studies that do exist in the health domain suffer methodological shortcomings, such as reliance on self-reported measures and non-random assignment to intervention groups (Hibbard et al., 2003, and Hibbard et al., 2005). The contribution of this paper is to disentangle the underlying mechanisms of private and public feedback incentives in the medical context in a controlled laboratory experiment. This method allows us to isolate the impact of feedback on the quality of medical care from other factors in the physician's decision environment, such as the simultaneous variation of financial incentives, regional system characteristics, and the health status of patients. Moreover, laboratory experiments are an inexpensive method to analyze the effects of a planned reform before it is implemented, and can thus help policy makers avoid costly failures. Specifically, our research adds to the literature discussed above in two main ways: Firstly, we investigate how non-monetary performance incentives for physicians affect the quality of the medical services they provide. Secondly, we control for the different motivation mechanisms behind public and private feedback by implementing the two separately and comparing their respective impact on the quality of medical care provided.

In our experiment subjects take on the role of physicians and make decisions over the medical treatment of patients, receiving feedback on the quality of care they provide. To account for the character of a political reform, we employ a within-subject design: In part 1 of the experiment subjects decide on the quantity of medical treatment they provide for a number of patients and are remunerated based on a fee-for-service schedule. In part 2 subjects are asked to make the same treatment decisions for an equal number of patients with the same characteristics as in part 1, but this time they will receive feedback on their performance at the end of the experiment in addition to the remuneration. Physician performance is measured in terms of outcome quality of care for the patient and is fully observable, i.e. not self-reported. Feedback is given in form of competitive rankings and is either private or public. Subjects who receive private feedback are informed about (only) their position in the ranking of participants on their computer screen. For public feedback subjects are asked to stand up while the ranking is read out loud by the experimenter, a procedure similar to that used in experimental studies by Rege and Telle (2004) and Ariely et al. (2009). In order to account for potential professional effects, we compare the decisions made by medical students – physicians in training – to those of other (student) subjects. Patient benefits realized in the experiment accrue to real patients as they are transferred to an organization which provides eye cataract operations.

In section 2 of this paper we describe our experimental design. In section 3 we present results, while section 4 discusses some policy implications and concludes.

2. Experimental Design

Our experiment consists of two parts, each containing a choice task with 9 decision situations. All subjects hence made a total of 18 individual decisions.

Decision Situations

The basic decision situation follows that of Brosig-Koch et al. (2013a, b).[1] The subject takes on the role of a physician and decides on the treatment of a patient. Treatment is performed by allotting the patient a quantity of $q\in[0,1,2, \ldots 10]$ medical services. With each treatment decision, the physician simultaneously determines his own profit $\pi(q)$ and the patient's health benefit $B_{kl}(q)$, both measured in monetary terms. For each treatment quantity, the physician also incurs costs c_{kl}=0.1q^2 which are deducted from his fee-for-service (FFS) remuneration R=2q.[2] This basic decision is repeated sequentially for nine patients, who differ in the benefit they stand to gain from medical treatment.[3] Each patient suffers from one of three illnesses, $k\in[A,B,C]$, which determines the maximum benefit he can receive from optimal treatment ($B_{Al}(q^*)$,= 7, $B_{Bl}(q^*)$,= 10, $B_{Cl}(q^*)$,= 14; see Figure 1). The illnesses each take on one of three degrees of severity, $l\in[x,y,z]$, which in turn determines the quantity of medical services at which a patient gains the optimal benefit from treatment (q_x^* = 3, q_y^* = 5, q_z^* = 7). See Appendix B.1 for a complete set of the parameters adapted from Brosig-Koch et al. (2013a).

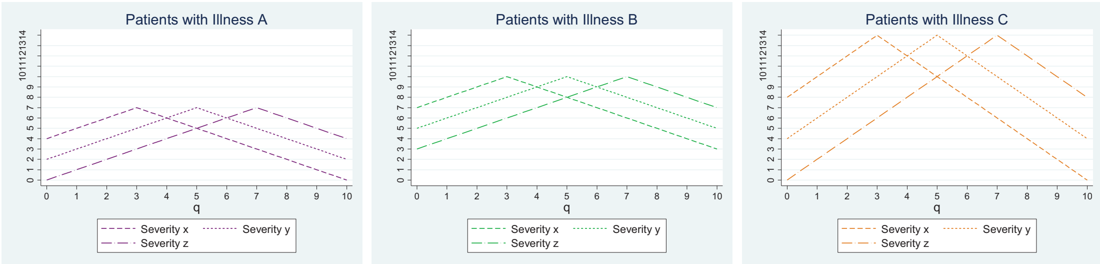
---

[1] Brosig-Koch et al. (2013b) study the effects of pay-for-performance incentives on physicians' provision behavior. Basing our experimental design on theirs allows us to compare financial and non-monetary incentive mechanisms in future research.

[2] We use FFS as it is the principal remuneration structure for primary physicians in most countries, e.g. in the US (Medicare), Australia, France, and Germany. Using a different payment structure such as capitation would presumably not change the qualitative results of our experiment, as we are concerned with a reform which is independent of monetary remuneration.

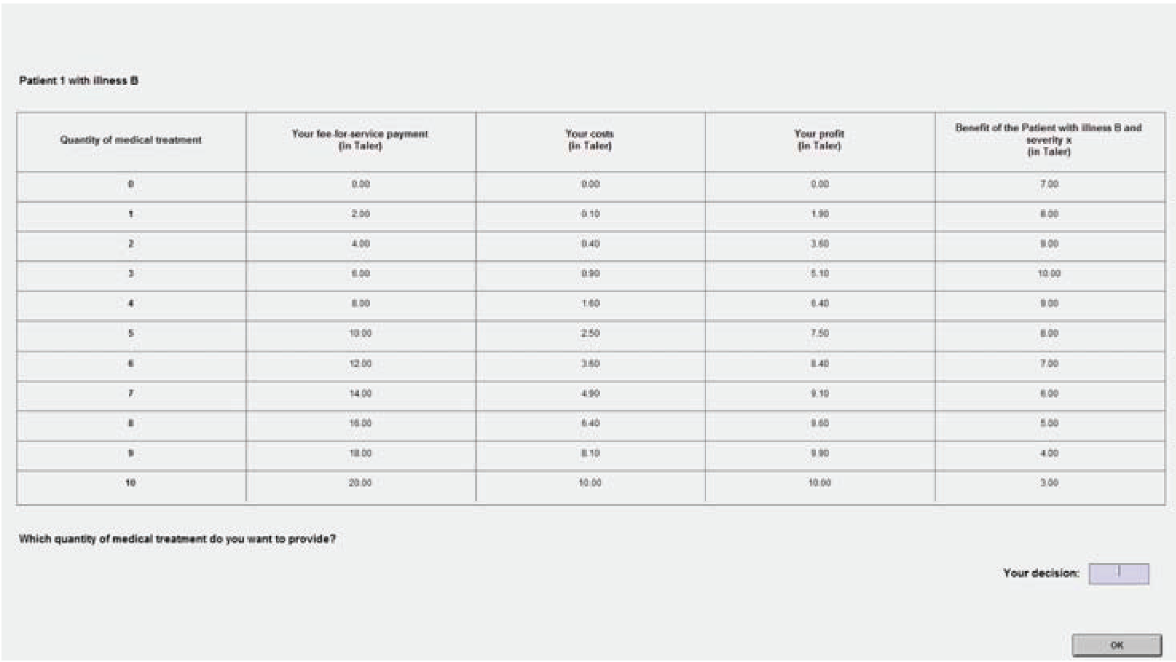
[3] The order of the 9 patients was determined randomly at the outset of the experiment and then kept constant for all subjects and in all variants of the choice task.

Figure 1: Patient benefit functions for illnesses k and severities l



The physician's profit-maximizing choice in every treatment decision is to provide the largest possible quantity of 10 medical services. As this quantity is always higher than the quantity that maximizes the patient's benefit (due to the fee-for-service remuneration scheme), subjects face a trade-off between the two welfare functions in each treatment decision. See Figure 2 for an example of the decision situation.

Figure 2: Example of a decision screen in treatments



Patient 1 with illness B

Quantity of medical treatment	Your fee-for-service payment (in Taler)	Your costs (in Taler)	Your profit (in Taler)	Benefit of the Patient with illness B and severity x (in Taler)
0	0.00	0.00	0.00	7.00
1	2.00	0.10	1.90	8.00
2	4.00	0.40	3.60	9.00
3	6.00	0.90	5.10	10.00
4	8.00	1.60	6.40	9.00
5	10.00	2.50	7.50	8.00
6	12.00	3.60	8.40	7.00
7	14.00	4.90	9.10	6.00
8	16.00	6.40	9.60	5.00
9	18.00	8.10	9.90	4.00
10	20.00	10.00	10.00	3.00

Which quantity of medical treatment do you want to provide?

Your decision:

OK

Patients

The patients in our experiment were not physically present in the laboratory. Nevertheless, the monetary value of the patient benefit went to real patients outside the laboratory. We follow Hennig-Schmidt et al. (2011) and Brosig et al. (2013a, b) in this approach to making patient outcomes in the decision situations directly relevant to health, rather than mere monetary payments. Subjects were instructed that the sum of all patient benefits achieved in the situations selected for payment would be transferred to the charity organization *Christoffel Blindenmission*, which provides care for patients with eye diseases.

Payment

All monetary amounts in the experiment were designated in the experimental currency of Taler. 1 Taler equals € 0.80. In keeping with experimental best practice, one decision situation for each part of the experiment was drawn at random at the end of the experiment (random payment technique) in order to avoid wealth and averaging effects.[4] The situations chosen in each session are valid for all its participants. Each subject received the combined physician profits achieved in these two situations as payoff for the experiment. The benefit received by the patient in these two situations was donated to the *Christoffel Blindenmission*. The donation was carried out immediately after the experimental session was completed and was witnessed by a randomly chosen subject (who received an additional payment of €5 for this task).

Treatment Conditions

In order to address our research questions, we conduct two separate treatment conditions: (1) PRIVATE and (2) PUBLIC.

In condition (1) PRIVATE, the first part of the experiment consists of the choice task as described above: subjects decide on medical treatment for 9 patients. In the second part of the experiment, subjects again make the same treatment decisions for these 9 patients. However, before beginning part two of the experiment, they are informed that at the end of this task, all participants in the session (typically 12 subjects) will be ranked according to the quality of treatment they provide. Treatment quality is defined as the (negative) difference between the realized patient benefit and the optimal patient benefit. The highest treatment quality is thus achieved by choosing the patient-optimal quantity of medical services; in this case treatment quality is zero. This performance feedback is given in private, so that subjects learn only their own position in the ranking (on their computer screen), but not anyone else's. Ranks are shared if participants provide equal treatment quality. Feedback is provided only for the one decision situation in this part of the experiment which has been randomly selected for payment.

Condition (2) PUBLIC is analogous to condition (1), consisting of the basic choice task in part one of the experiment and a feedback incentive for the choice task in part two. Again, subjects are told in the instructions for part two that they will be ranked according to the quality of treatment provided in the situation chosen for payment. In this condition, however, the ranking is made public among the participants of this session: First, the rank table with all participants (identified by their seat numbers) is displayed on their computer screens (see Figure 3). Next, in a procedure similar to that of Ariely et al. (2009) and Rege and Telle (2004), subjects are requested to stand up (allowing everyone to see everyone else over the walls of their cubicles). The ranking is then read aloud by the

[4] Various studies confirm that the random payment technique does not dilute the power of the monetary incentive for non-complex choice tasks (Starmer and Sugden, 1991, Cubitt et al. 1998, Laury, 2006, Baltussen et al., 2010).

experimenter. As they are called up, subjects are required to raise a sign displaying their seat number.

We use a within-subject design to account for the character of a reform that introduces performance feedback. This allows us to analyze behavior before and after the reform in a controlled way: Comparing the decisions made in part one in these two treatments to those made in the incentivized tasks in part two (within-subject comparison) permits us to address our research question Q1 whether feedback incentives have an impact on the quality of medical treatment provided. The comparison of choices made in part two between treatments (1) and (2) helps us answer our research question Q2 whether the mode of delivering feedback – privately or publicly – affects the impact of the feedback incentive on treatment quality.

Figure 3: Example public feedback screen

Medical Students

In all sessions of our experiment, we recruited medical students as well as students of other degree programs as subjects. Comparing decision behavior between these groups allows us to clarify whether prospective physicians – who have perhaps selected themselves into medical education based on specific social preferences, or are influenced by medical professional norms in the course of their training, or both – react differently to reputation-based performance incentives. Ahlert et al. (2012), for example, find that behavior in situations framed as medical treatment decisions (rather than neutral decisions) is impacted by the professional norms of medicine or economics adopted by their subjects. However, other experiments carried out at the Essen Laboratory for Experimental Economics involving different types of health-related decisions have not confirmed this type of professional effect (e.g. Brosig-Koch et al. 2013a, b).

Robustness Check

We test the robustness of our results against the order in which subjects face the incentivized and non-incentivized tasks. Aside from experimental design considerations, private feedback could have motivating or demotivating effects on provision behavior in the second part of the experiment. We reversed the task order in two sessions for treatment condition (1) PRIVATE FEEDBACK: Subjects here completed part 1 with a private feedback incentive and part 2 without a feedback incentive. Note that we could not test for a reverse task order with public feedback as this would imply the loss of subjects' anonymity in part 1 of the experiment, which compromises subsequent decisions in the non-incentivized task in part two of the experiment.

Experimental Procedure

The experiment was carried out at the Essen Laboratory for Experimental Economics (Duisburg-Essen University) in June 2012 using the specialized software z-tree (Fischbacher 2007). 144 subjects were recruited via ORSEE (Greiner, 2004) and participated in a total of 12 sessions of about an hour each.

Subjects were allocated to seats in the laboratory by a random draw. They received separate written instructions at the outset of each part of the experiment and were given several minutes to read the instructions carefully and to ask clarifying questions. At the beginning of part 1, subjects also completed several control questions (see Appendix A) which served to ensure that all subjects understood the task at hand. The control questions were announced in the instructions and were not relevant to any payments earned in later decisions.

At the end of the experiment all subjects were paid out individually and in private. They received an average payoff of €13.51 (min: €7.6, max: €16.00) and generated an average patient benefit of €12.18 (min: € 2.4, max: € 22.4). In total, €1754.4 were transferred to the *Christoffel-Blindenmission*. Assuming a cost of €30 per eye cataract operation, this amounts to the treatment of about 58 real patients.

3. Results

Data

We consider decisions made by 144 subjects. See Table 1 for the distribution of participants across treatment conditions and degree programs.

Table 1: Overview subjects

Treatment	Number of subjects		
	Total	**Medical students**	**Others**
(1) PRIVATE	60	12	48
(2) PUBLIC	60	14	46
(3) REVERSE ORDER (PRIVATE)	24	5	19
Total	144	31	113

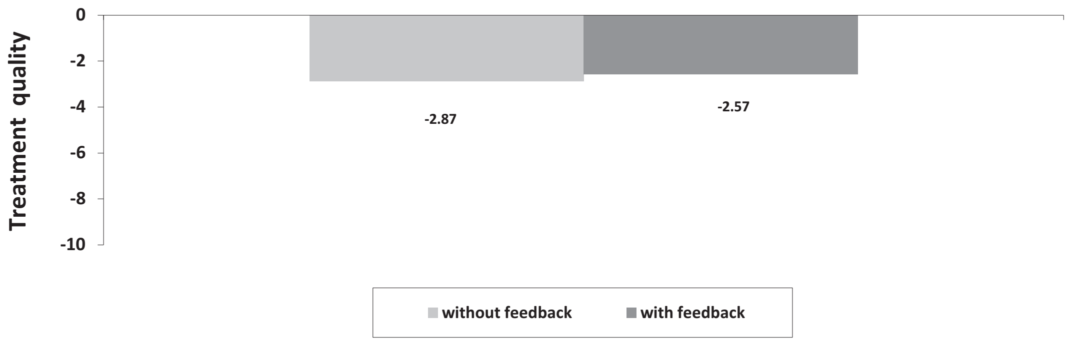
Impact of Feedback Incentives on Treatment Quality

In order to analyze whether feedback incentives serve to improve the quality of medical treatment provided, we first consider the decisions made by all subjects in treatment conditions (1) and (2) and compare their choices in the first task without a feedback incentive to those in the second task with feedback. Treatment quality is defined as the (negative) difference between the optimal benefit a patient can potentially achieve from being treated and the actual benefit he receives from the amount of services he is provided. Average treatment quality thus ranges from 0 (no deviation from optimal quality) to -10.3 (the largest possible average deviation from the optimum across all 9 decisions).

We consider the aggregated decisions made by our subjects for all patients and across all illnesses and degrees of severity, as this best reflects the typical decision situation of a physician who is faced with a heterogeneous group of patients within a time interval such as a month or a quarter.[5]

[5] We control for the impact of the individual illnesses and degrees of severity on the physician's treatment quality in an OLS regression and find significant coefficients for both (see Appendix B.2). This does not detract from our results, as our main concern in this paper is with the general situation of a physician facing a heterogeneous group of patients. However, the impact of feedback incentives on the performance of physicians who deal with more specific sub-populations of patients (e.g. with particularly severe or chronic illnesses) is an interesting subject of further research.

Figure 4: Average treatment quality by task



In the aggregate, the subjects in our experiment provided treatment with a quality of -2.87 on average (so their decisions result in an average loss of patient benefit of 2.87 Taler relative to the optimum; SD = 2.86) in decisions without a feedback incentive, and of -2.57 (SD = 2.74) in decisions with feedback (see Figure 4). This difference is highly statistically significant in a two-sided Mann-Whitney U-test ($p < 0.01$).[6] Our first result is thus:

> *In general, setting a non-monetary feedback incentive for subjects significantly improves the quality of medical treatment they provide to patients.*

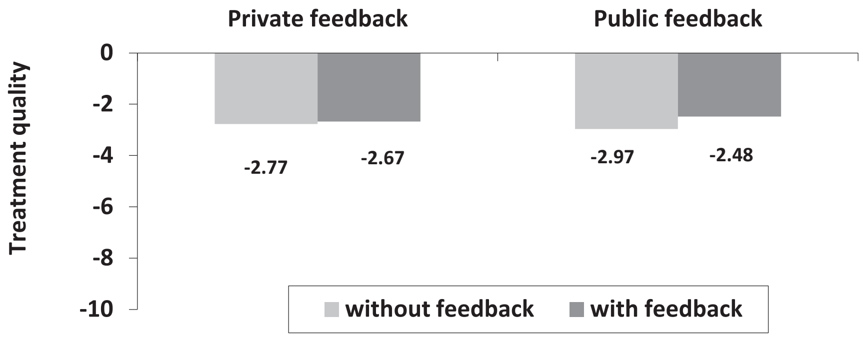
Effect of Feedback Mode

Turning to the relative effects of giving performance feedback privately or publicly, we compare the effect of the feedback incentive across the treatment conditions PUBLIC and PRIVATE. The public feedback incentive in treatment condition (2) led to an improvement in the medical treatment quality from -2.97 (SD = 2.83) to -2.48 (SD = 2.68; see Figure 5).[7] This difference is statistically highly significant ($p < 0.01$). In treatment condition (1), the private feedback incentive improved the average treatment quality slightly from -2.77 (SD = 2.88) to -2.67 (SD = 2.79). This shift is, however, not statistically significant ($p = 0.64$). (The results of these statistical tests are also confirmed in simple OLS regressions; see Appendix B.3.)

[6] Unless noted otherwise, all statistical tests presented here are two-sided Mann-Whitney U-tests and two-tailed Student's t-tests provide very similar results.

[7] Note that while subject behavior in task 1 differs slightly across treatments (1) and (2), this difference is not statistically significant ($p > 0.10$).

Figure 5: Average treatment quality by feedback mode



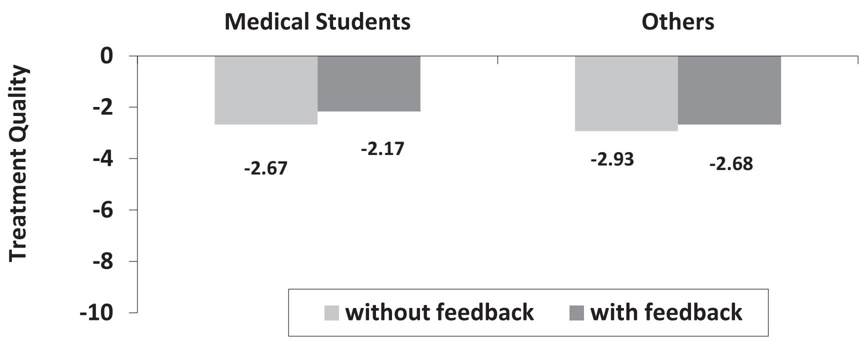
This leads to our second result:

> *The mode in which feedback incentives are provided matters: While public feedback yields a significant improvement in the treatment quality subjects provide, the effect of private feedback is not statistically significant.*

Medical Students

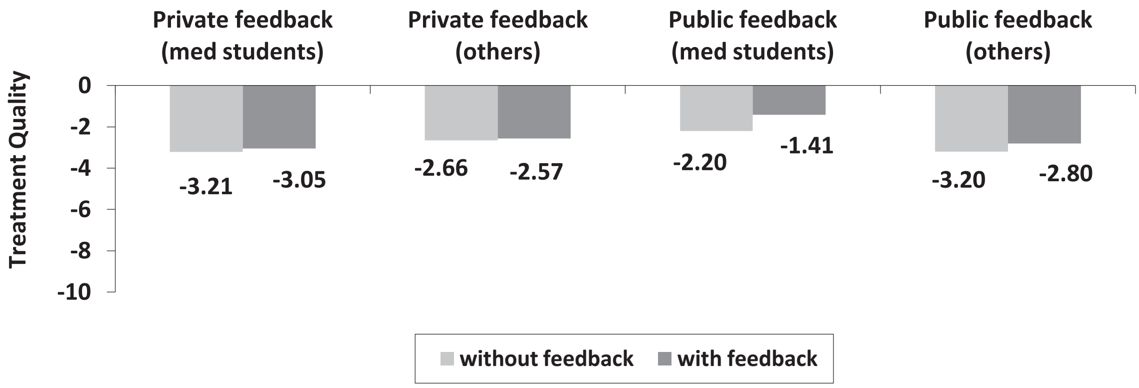
The above two results are generally robust to a relevant subject pool characteristic, whether subjects medical students or not. Considering sub-samples of medical students and other subjects separately, feedback incentives improve average treatment quality from -2.67 (SD = 2.48) to 2.17 (SD = 2.34) for the prior and from -2.93 (SD = 2.95) to -2.68 (SD = 2.83) for the latter (see Figure 6). Both shifts are statistically significant: p = 0.01 and p = 0.07, respectively.

Figure 6: Average treatment quality by degree



The impact of the feedback mode also holds for the two separate sub-samples (see Figure 7): Private feedback tends to improve treatment quality, though the effect is not statistically significant: Medical students in this group achieve a quality of -3.21 (SD = 2.81) without and -3.05 (SD = 2.66) with the incentive, while other subjects improve very slightly from -2.66 (SD = 2.89) to -2.57 (SD = 2.82). For both subsamples, the differences are not statistically significant (p > 0.70). The significant effect of the public feedback incentive, on the other hand, is upheld in both groups: Medical students improve their treatment quality from -2.20 (SD = 2.06) to -1.41 (SD = 1.71), while others improve from -3.20 (SD = 2.99) to -2.80 (SD = 2.83); both changes are statistically significant, with p < 0.01 and p < 0.05, respectively.[8] Moreover, simple OLS regressions show that given public feedback, medical students provide significantly better treatment quality than non-medical students (see Appendix B.4).

Figure 7: Average treatment quality by feedback mode and degree



Hence we find that public feedback significantly improves the treatment quality provided by medical and other students, while the effect of private feedback is not statistically significant for both groups. The effect for public feedback is significantly larger for medical students.

Robustness to Task Order and Subject Characteristics

Using data from the two reverse-order sessions, we find that the results of the private feedback incentive are robust to providing the quality incentive in part 1 of the experiment and not providing it in part 2. Subjects achieve an average treatment quality of -2.38 (SD = 2.52) and -2.13 (SD = 2.45) respectively, which does not represent a statistically significant difference (p = 0.31; see Figure 8).

As mentioned above, corresponding controls for a reversed task order are difficult to implement in the public feedback treatment. Making subjects' decisions or their consequences known to other participants in the experiment in part 1 would presumably have an additional influence on the decisions made in part 2, obfuscating the effect of purely reversing the tasks.

[8] While the within-subject effect of public and private feedback is consistent across groups, the treatment quality provided in part one (the non-incentivized task) differs significantly across all pairs of subject groups discussed in this section (p<0.05).

We also estimated OLS regressions to control for the influence of subject characteristics (age, gender, family members in the medical profession) and specifics of the decision situation (severity, illness, session, whether subjects knew other participants in the session) on the quality of medical treatment provided in our experiment. None of these factors adds any explanatory power to our analysis (see Appendix B.4).

4. Conclusion

We find feedback as a performance incentive to have an effect on the quality of medical care provided in our experiment. The effect is, however, dependent on the feedback mode: Private feedback has no impact on the quality of care, whereas public feedback has a significant positive impact. Our results are robust to a subject's enrollment in medical education and socio-demographic characteristics as well as to changes in the task order.

So far, there is evidence that physicians react to non-monetary performance incentives (Kolstad, 2013), However, there seems to be little evidence that quality reporting incentives actually lead to better medical treatment quality and lower health care costs (Dranove and Jin, 2010). This may be due to the fact that while treatment quality is typically multidimensional, only some of its aspects can be reported, as e.g. in the US Nursing Home Quality Initiative. In this case, physicians may react to public reporting by improving quality only for the reported measures whilst decreasing quality along non-reported dimensions, for instance by patient selection (Dranove et al., 2003, Werner and Asch, 2005). In our controlled laboratory experiment quality is fully reported. Under these circumstances, we find public feedback incentives to have a positive and significant effect on the quality of medical care provided. Hence, if future policy reforms succeed at establishing more comprehensive ways of reporting quality in health care, this should serve as a tool to increase quality of care.

Our results also suggest that the mode of providing quality feedback is important and should be taken into account by policy makers. The mere motive of boosting self-esteem which underlies private performance feedback does not seem sufficient to align physician interests more closely with patient interests. The additional motive of reputation (image motivation) introduced by public performance feedback, on the other hand, can perhaps foster quality improvement in medical care. Public performance feedback may be a cost-efficient means towards this end – in contrast to monetary pay-for-performance incentives, which also serve to raise patient benefit but are not necessarily cost-efficient (Brosig-Koch et al., 2013b). Future research in this area should be directed towards investigating how monetary mechanisms interact with non-monetary mechanisms, and the conditions under which they enhance or detract from each other.

Literature

Ahlert, M., Felder, S., and B. Vogt (2012): Which Patients Do I Treat?: An Experimental Study with Economists and Physicians, *Health Economics Review*, 2(1), 1–11.

Ariely, D., A. Bracha, S. Meier (2009): Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially, *American Economic Review*, 99, 544–55.

Azmat, G., and N. Iriberri (2010): The importance of relative performance feedback information: Evidence from a natural experiment using high school students, *Journal of Public Economics*, 94(7-8), 435-452.

Baltussen, G., Post, T., van den Assem, M. & Wakker, P.P. (2010): Random Incentive Systems in a Dynamic Choice Experiment, Working Paper, Erasmus University of Rotterdam.

Barankay, I., (2011a): Rankings and social tournaments: evidence from a crowd-sourcing experiment, Working Paper, Wharton School of Business, University of Pennsylvania.

Barankay, I., (2011b): Gender differences in productivity responses to performance rankings: evidence from a randomized workplace experiment, Working Paper. Wharton School of Business, University of Pennsylvania.

Blanes i Vidal, J. and M. Nossol (2011): Tournaments without prizes: evidence from personnel records, *Management Science*, 57 (10), 1721-1736.

Bénabou, R., J. Tirole (2002): Self-confidence and personal motivation, *Quarterly Journal of Economics*, 117(3), 871–915.

Bénabou, R., J. Tirole (2006): Incentives and Prosocial behaviour, *The American Economic Review*, 96, 1652-1678.

Brosig-Koch, J., H. Hennig-Schmidt, N. Kairies, and D. Wiesen (2013a): How to improve patient care? An analysis of capitation, fee-for-service, and mixed incentive schemes for physicians, Ruhr Economic Papers No. 412.

Brosig-Koch, J., H. Hennig-Schmidt, N. Kairies, and D. Wiesen (2013b): How effective are pay-for-performance incentives for physicians? A laboratory experiment, Ruhr Economic Papers No. 413.

Charness, G., Masclet, D. and M.C. Villeval (2011): Competitive preferences and status as an incentive: experimental evidence. CIRANO Working Paper.

Cubitt, R., Starmer, R., Sugden, R. (1998): On the validity of the random lottery incentive system, Experimental Economics, 1(2), 115-131.

Cutler, D.M., Ilckman, R.S. and M. B. Landrum (2004): The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery, *American Economic Review*, 94(2), 342-346.

Dranove, D., Kessler, D., McClellan, M. and M. Satterthwaite (2003): Is More Information Better? The Effects of "Report Cards" on Health Care Providers, *Journal of Political Economy*, 111(3), 555-588.

Dranove, D., and G. Zhe Jin (2010): Quality Disclosure and Certification: Theory and Practice, *Journal of Economic Literature*, 48(4), 935–63.

Ellingsen, T., M. Johannesson (2008): Pride and Prejudice: the Human Side of Incentive Theory, *American Economic Review*, 98, 990-1008.

Eriksson, T., A. Poulsen, M.C. Villeval (2009): Feedback and Incentives: Experimental Evidence, *Labour Economics*, 16, 679-688.

Fischbacher, U. (2007): Z-tree: Zurich toolbox for readymade economic experiments, Experimental Economics, 10(2), 171-178.

Grabowski, D.C., and Town, R.J., (2011): Does information matter? Competition, quality, and the impact of nursing home report cards, *Health Services Research*, 46, 1698–1719.

Greiner, B. (2004): The Online Recruitment System ORSEE 2.0 - A Guide for the Organization of Experiments in Economics, University of Cologne working paper No. 10.

Hannan, E. L., Kilburn, H., Racz, M., Shields, E. and M. R. Chassin (1994): Improving the Outcomes of Coronary Artery Bypass Surgery in New York State, *Journal of the American Medical Association*, 271(10), 761–66.

Hannan, R.L., Krishnan, R., Newman, A. (2008): The effects of disseminating relative performance feedback in tournament and individual performance compensation systems, *Accounts Review*, 83, 893–913.

Hennig-Schmidt, H., R. Selten and D. Wiesen (2011): How Payment Systems Affect Physicians' Provision Behavior? An Experimental Investigation, *Journal of Health Economics*, 30(4), 637-646.

Hibbard, J., Stockard, J. and M. Tusler (2003): Does Publicizing Hospital Performance Stimulate Quality Improvement Efforts?, *Health Affairs*, 22(2), 84–94.

Hibbard, J.H., Stockard, J., and M. Tusler (2005): Hospital performance reports: impact on quality, market share and reputation. *Health Affairs*, 24(4), 1150–1160.

Kolstad, J. T. (2013): Information and quality when motivation is intrinsic: Evidence from surgeon report cards, National Bureau of Economic Research No. w18804.

Kuhnen, C., and A. Tymula (2012): Feedback, self-esteem and performance in organizations, *Management Science,* 58, 94–113.

Laury, S.K. (2006): Pay One or Pay All: Random Selection of One Choice for Payment, Working Paper, Georgia State University.

Mas, A., and E. Moretti (2009): Peers at work, *American Economic Review*, 99(1), 112–145.

Marshall, M.N., Shekelle, P.G., Leatherman, S., and R.H. Brook (2000): The Public Release of Performance Data: What Do We Expect to Gain? A Review of the Evidence, *Journal of the American Medical Association,* 1866–1874.

Maynard, A. (2012): The powers and pitfalls of payment for performance, *Health Economics*, 21 (1), 3-12.

McClellan, M. (2011): Reforming payments to healthcare providers: The key to slowing healthcare cost growth while improving quality?, Journal of Economic Perspectives, 25, 69-92.

Rege, M. and K. Telle (2004): The Impact of Social Approval and Framing on Cooperation in Public Good Situations, *Journal of Public Economics*, 88, 1625-1644.

Rosenthal, G. E., Quinn, L., and D. L. Harper (1997): Declines in hospital mortality associated with a regional initiative to measure hospital performance, *American Journal of Medical Quality*, 12(2), 103-112.

Starmer, C. and R. Sugden (1991): Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation, American Economic Review, 81(4): 971-978.

Tran, A., and R. Zeckhauser (2012): Rank as an inherent incentive: Evidence from a field experiment, *Journal of Public Economics*, 96(9-10), 645–650.

Werner, R.M., and D. A. Asch (2005): The Unintended Consequences of Publicly Reporting Quality Information, *Journal of the American Medical Association*, 293(10), 1239-1244.

Werner, R.M., Konetzka, R.T., Stuart, E.A., Norton, E.C., Polsky, D., and J. Park (2009): The impact of public reporting on quality of postacute care, *Health Services Research*, 44(4), 1169–87.

Appendix

Appendix A: Instructions and Control Questions

You are participating in an economic experiment on decision behavior. You and the other participants will be asked to make decisions for which you can earn money. Your payoff depends on the decisions you make. At the end of the experiment, your payoff will be converted to Euro and paid to you in cash. During the experiment, all amounts are presented in the experimental currency Taler. 10 Taler equals 8 Euro.

The experiment will take about 90 minutes and consists of two parts. You will receive detailed instructions before each part. Note that none of your decisions in either part have any influence on the other part of the experiment.

Part One

Please read the following instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you.

Part one of the experiment consists of 9 rounds of decision situations.

Decision Situations

In each round you take on the role of a physician and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness.

Every patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Profit

In each round you are remunerated for treating the patient. Your remuneration increases with the amount of medical treatment you provide. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the remuneration.

Every quantity of medical service yields a particular benefit for the patient – contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit but also the patient's benefit.

In each round you will receive detailed information on your screen (see below) on the patient's illness and its severity as well as the remuneration, cost, and patient benefit for each quantity of medical services (see screen shot in Figure 1 above).

Payment

At the end of the experiment one of the 9 rounds of part one will be chosen at random. Your profit in this round will be paid to you in cash.

For this part of the experiment, no patients are physically present in the laboratory. Yet, the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, an organization which funds the treatment of patients with eye cataract.

The transfer of money to the Christoffel Blindenmission Deutschland e.V. will be carried out after the experiment by the experimenter and one participant. The participant completes a money transfer form, filling in the total patient benefit (in Euro) resulting from the decisions made by all participants in the randomly chosen situation. This form prompts the payment of the designated amount to the Christoffel Blindenmission Deutschland e.V. by the University of Duisburg-Essen's finance department. The form is then sealed in a postpaid envelope and posted in the nearest mailbox by the participant and the experimenter.

After the entire experiment is completed, one participant is chosen at random to oversee the money transfer to the Christoffel Blindenmission Deutschland e.V. The participant receives an additional compensation of 5 Euro for this task. The participant certifies that the process has been completed as described here by signing a statement which can be inspected by all participants at the office of the Chair of Quantitative Economic Policy. A receipt of the bank transfer to the Christoffel Blindenmission Deutschland e.V. may also be viewed here.

Comprehension Questions

Prior to the decision rounds we kindly ask you to answer a few comprehension questions. They are intended to help you familiarize yourself with the decision situations. If you have any questions about this, please raise your hand. Part one one if the experiment will begin once all participants have answered the comprehension questions correctly.

Part Two

Please read the following instructions carefully. We will approach you in about five minutes to answer any questions you may have. If you have questions at any time during the experiment, please raise your hand and we will come to you.

Part two of the experiment also consists of 9 rounds of decision situations.

Decision Situations

As in part one of the experiment, you take on the role of a physician in each round and decide on medical treatment for a patient. That is, you determine the quantity of medical services you wish to provide to the patient for a given illness and a given severity of this illness.

Every patient is characterized by one of three illnesses (A, B, C), each of which can occur in three different degrees of severity (x, y, z). In each consecutive decision round you will face one patient who is characterized by one of the 9 possible combinations of illnesses and degrees of severity (in random order). Your decision is to provide each of these 9 patients with a quantity of 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Profit

In each round you are remunerated for treating the patient. Your remuneration increases with the amount of medical treatment you provide. You also incur costs for treating the patient, which likewise depend on the quantity of services you provide. Your profit for each decision is calculated by subtracting these costs from the remuneration.

As in part one, every quantity of medical service yields a particular benefit for the patient – contingent on his illness and severity. Hence, in choosing the medical services you provide, you determine not only your own profit but also the patient’s benefit.

In each round you will receive detailed information on your screen (see below) on the patient’s illness and its severity as well as the remuneration, cost, and patient benefit for each quantity of medical services (see screen shot below).

Payment

At the end of the experiment one of the 9 rounds of part two will be chosen at random. Your profit in this round will be paid to you in cash, in addition to your payment from the round chosen for part one of the experiment.

After the experiment is over, please remain seated until the experimenter asks you to step forward. You will receive your payment at the front of the laboratory before exiting the room.

As in part one, no patients are physically present in the laboratory for part two of the experiment. Yet, the patient benefit does accrue to a real patient: The amount resulting from your decision will be transferred to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, an organization which funds the treatment of patients with eye cataract.

The process for the transfer of money to the Christoffel Blindenmission Deutschland e.V. as described for part one of the experiment will be carried out by the experimenter and one participant.

<u>Feedback</u>

In addition to your payment you will receive feedback in this part of the experiment on the quality of treatment you provide as a physician. The best treatment quality is achieved when the patient receives the highest possible benefit. The lower the patient's benefit from the provided amount of services, the worse the treatment quality.

A ranking of all participants in the experiment will be generated. The ranking is based on the treatment quality provided in the decision situation chosen for payment in this part of the experiment. The participant with the highest treatment quality ranks first, the participant with the worst treatment quality ranks last. Participants with equal treatment quality share ranks.

[Private feedback treatment:]

You will see your placement in this ranking on your screen at the end of the experiment. Every participant only learns their own rank, not those of other participants.

[Public feedback treatment:]

This ranking will be shown on your screen once the experiment has been completed. A member of the laboratory staff will then ask all participants to stand up. The ranking will be read out aloud. (The participants' ranks and seat number will be stated, not their names or specific decisions.) When your seat number is called, please hold up the sign with the number so that it is visible to all participants.

We kindly ask you to not talk to anyone about the content of this session in order to prevent influencing other participants after you. Thank you for your Collaboration!

Exemplary Comprehension Question Part 1:

Quantity of medical treatment	Fee-for-service (in Taler)	Costs (in Taler)	Profit (in Taler)	Benefit of the patient with illness F and severity y (in Taler)
0	0.00	0.00	0.00	15.00
1	4.00	0.20	3.80	16.00
2	8.00	0.80	7.20	17.00
3	12.00	1.80	10.20	18.00
4	16.00	3.20	12.80	19.00
5	20.00	5.00	15.00	20.00
6	24.00	7.20	16.80	19.00
7	28.00	9.80	18.20	18.00
8	32.00	12.80	19.20	17.00
9	36.00	16.20	19.80	16.00
10	40.00	20.00	20.00	15.00

Assume that a physician wants to provide 2 quantities of medical treatment for the patient depicted above.

a) What is the fee-for-service?
b) What are the costs?
c) What is the profit?
d) What is the patient benefit?

Appendix B: Further Tables

B.1 Decision Parameters

		Quantity (q)										
Treatment	Variable	0	1	2	3	4	5	6	7	8	9	10
all	$R_{kl}^{Part\ 1}$	0	2	4	6	8	10	12	14	16	18	20
	$R_{kl}^{Part\ 2}$	0	2	4	6	8	10	12	14	16	18	20
all	c_{kl}	0	0.1	0.4	0.9	1.6	2.5	3.6	4.9	6.4	8.1	10
all	π_{kl}^{Part1}	10	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6	1.9	0
	π_{kl}^{Part2}	10	9.9	9.6	9.1	8.4	7.5	6.4	5.1	3.6	1.9	0
all	B_{Ax}	4	5	6	7	6	5	4	3	2	1	0
	B_{Ay}	2	3	4	5	6	7	6	5	4	3	2
	B_{Az}	0	1	2	3	4	5	6	7	6	5	4
	B_{Bx}	7	8	9	10	9	8	7	6	5	4	3
	B_{By}	5	6	7	8	9	10	9	8	7	6	5
	B_{Bz}	3	4	5	6	7	8	9	10	9	8	7
	B_{Cx}	8	10	12	14	12	10	8	6	4	2	0
	B_{Cy}	4	6	8	10	12	14	12	10	8	6	4
	B_{Cz}	0	2	4	6	8	10	12	14	12	10	8

B.2 Control for Illnesses and Severities (OLS regression, aggregated data)

VARIABLES	Aggregated
Severity	0.647***
	(0.039)
Illness	-1.013***
	(-0.0737)
Constant	-3.928***
	(0.241)
Observations	2,160
R-squared	0.229
N_clust	120

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

B.3 Effect of Feedback Mode (OLS regression)

VARIABLES	Private	Public
Feedback Incentive	0.106	0.491***
	(0.0808)	(-0.0989)
Constant	-2.774***	-2.967***
	(0.23)	(0.253)
Observations	1,080	1,080
R-squared	0.000	0.008
N_clust	60	60

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

B.4 Control for Subject Characteristics

B.4.1 Descriptive Subject Characteristics (excluding reverse-order subjects):

Variable	**Average (120 subjects)**	**Min**	**Max**
Female	0.6	0	1
Age	23.5	18	49
Medical student	0.142	0	1
Econ student	0.325	0	1
Parents in health care job	0.2	0	1
Knew other subjects in session	0.31	0	3

B.4.2 Regressions (OLS, clustered by subjects, excluding reverse-order subjects):

Aggregate Data

VARIABLES	1	2	3	4	5
Feedback Incentive	0.298***	0.298***	0.298***	0.298***	0.298***
	(0.066)	(0.066)	(0.066)	(0.066)	(0.066)
Age	0.00729				
	(0.0467)				
Female		0.41			
		(0.357)			
Parents in health care job			-0.291		
			(0.463)		
Medical student				0.389	
				(0.371)	
Knew other subjects in session					-0.434
					(0.342)
Constant	-3.042***	-3.117***	-2.812***	-2.955***	-2.736***
	(1.091)	(0.285)	(0.186)	(0.201)	(0.18)
Observations	2,160	2,160	2,160	2,160	2,160
R-squared	0.003	0.008	0.005	0.006	0.011
N_clust	120	120	120	120	120

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Private Feedback

VARIABLES	1	2	3	4	5
Feedback Incentive	0.106	0.106	0.106	0.106	0.106
	(0.0808)	(0.0808)	(0.0808)	-0.0808	(0.0808)
Age	0.134*				
	(0.0703)				
Female		0.0243			
		(0.474)			
Parents in health care job			(0.849)		
			(0.537)		
Medical student				0.51	
				(0.586)	
Knew other subjects in session					-0.940*
					(0.514)
Constant	-5.819***	-2.788***	-2.533***	-2.672***	-2.524***
	(1.635)	(0.346)	(0.258)	(0.257)	(0.227)
Observations	1,080	1,080	1,080	1,080	1,080
R-squared	0.023	0	0.019	0.006	0.025
N_clust	60	60	60	60	60

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Public Feedback

VARIABLES	1	2	3	4	5
Feedback Incentive	0.491***	0.491***	0.491***	0.491***	0.491***
	(0.099)	(0.099)	(0.099)	(0.099)	(0.099)
Age	-0.0322				
	(0.0488)				
Female		0.823			
		(0.548)			
Parents in health care job			0.772		
			(0.859)		
Medical student				1.194***	
				(0.415)	
Knew other subjects in session					-0.168
					(0.465)
Constant	-2.185*	-3.488***	-3.057***	-3.245***	-2.908***
	(1.174)	(0.465)	(0.264)	(0.308)	(0.278)
Observations	1,080	1,080	1,080	1,080	1,080
R-squared	0.012	0.028	0.016	0.041	0.009
N_clust	60	60	60	60	60

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Chapter 5

P4P, Reputation and the Reduction of Costly Overprovision

Reference

Kairies, N. (2012): “P4P, Reputation and the Reduction of Costly Overprovision.” *Ruhr Economic Papers* No. 331.

RUHR
ECONOMIC PAPERS

Nadja Kairies

P4P, Reputation and the Reduction of Costly Overprovision

UNIVERSITÄT DUISBURG ESSEN

#331

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Editorial Office

Ruhr Economic Papers #331

Responsible Editor: Volker Clausen

ISSN 1864-4872 (online) – ISBN 978-3-86788-380-1

Ruhr Economic Papers #331

Nadja Kairies

P4P, Reputation and the Reduction of Costly Overprovision

UNIVERSITÄT DUISBURG ESSEN

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

http://dx.doi.org/10.4419/86788380
ISSN 1864-4872 (online)
ISBN 978-3-86788-380-1

Nadja Kairies[1]

P4P, Reputation and the Reduction of Costly Overprovision

Abstract

We investigate the effect of reputational motivation on output in a scenario of overprovision of medical treatment. We assume that physicians differ in their degree of altruism, enjoy being perceived as good but dislike being perceived as greedy. We show that better reputational motivation unambiguously reduces the costs of healthcare provision and the magnitude of overprovision which in turn rises patient benefits.

JEL Classification: D64, I11, I18, H42

Keywords: Altruism; performance; motivation

April 2012

1 University of Duisburg-Essen. – All correspondence to: Nadja Kairies, Chair of Quantitative Economic Policy, University of Duisburg-Essen, Faculty of Economics, Universitätsstr. 12, 45117 Essen, Germany, Email: nadja.kairies@ibes.uni-due.de.

1 Introduction

For policy makers in the healthcare sector improving the quality of care and reducing costs have been of major interest in recent years, see McClellan (2011). In particular, in case of overprovision both aims are not mutually exclusive as a reduction of medical treatment may increase patients' health benefits and decrease costs for the healthcare payer, see Cutler and Ly (2011). In practice, this is likely if the fee-for-service (FFS) is high, see e.g. Medicare in the US. Recent reforms targeting theses issues make use of financial incentives for predetermined performance measures (pay-for-performance), see e.g. Rosenthal et al. (2004) for the US and Doran et al. (2006) for the UK, or non-monetary incentives that strengthen physicians' reputational motivation by improving public quality reporting, see e.g. Glazer et al. (2007). The latter reforms are based on the empirical evidence that physicians actually care about their reputation and want to appear as "good" physicians.[1]

In the light of those reforms and along the lines of Siciliani (2009) we provide a theoretical framework with both monetary and non-monetary incentives for physicians. In contrast to Siciliani (2009) we focus on overprovision and reputational motivation. The monetary incentive is modeled by a FFS for each quantity of care. If we abstract from reputational motivation a higher FFS yields an increase in care since it increases physicians' marginal revenues. Concerning the reputational motivation we introduce a patient benefit function which links the quantity of care and the corresponding patient benefit. We assume that physicians are altruistic like in Ellis and McGuire (1986) and care about their reputation like in Bénabou and Tirole (2006). As in Ellis and McGuire (1986) but in contrast to Siciliani (2009) we assume that patients are characterized by a "peaked" patient benefit function which allows for both under- and overprovision of care. In case under- or overprovision is moderate physicians receive an extra utility gain since they are perceived as "good" type. This reputational motivation is different across physicians since they are heterogeneous in their degree of altruism. Physicians who are more altruistic adjust the quantity of care in order to be perceived as "good" while low altruism physicians maximize profits.

We use the model framework to derive comparative static results with respect to the total amount of care. We show that better reputational motivation unambiguously reduces the magnitude of overprovision which in turn increases patient benefits and decreases costs of healthcare provision. We then introduce a measure for the efficiency of the FFS scheme.

[1]Dranove et al. (2003) and Werner and Asch (2005) show that the introduction of report cards changed physicians' behavior in order to be classified as "good".

Intuitively, an efficiency maximizing FFS trades off the physicians' demand for a high FFS (higher marginal revenues) with the patients' aim for a low FFS (less overprovision). We show that an efficiency maximizing FFS exists and decreases if reputational motivation increases. For policy makers this can be an important result since promoting reputational motivation may actually increase patient benefits and simultaneously decrease costs.

2 Model

Let q denote the quantity of medical treatment that a physician provides to a patient.[2] Physicians differ in their degree of altruism θ as in Ellis and McGuire (1986), where θ can take a continuum of values $\theta \in [\underline{\theta}, \overline{\theta}]$.[3] The corresponding density function is denoted by $f(\theta)$ while the cumulative density function is $F(\theta)$. For a physician the provision of q involves total costs $C(q)$ with $C_q > 0$ and $C_{qq} > 0$. This cost function C includes all the monetary and non-monetary costs associated with the provision of q.

A patient benefits from the provision of medical treatment q. Formally, the patient benefit is denoted by $B(q)$ and we assume that there exists a unique global maximum $B(q_B^*)$. Economically, if a physician provides $q < q_B^*$ ($q > q_B^*$) the patient suffers from underprovision (overprovision).

A physician's remuneration is composed of a capitation $T > 0$ and a FFS $p > 0$, which are both financed by the healthcare payer. Hence, profits are

$$\pi = T + pq - C(q) . \tag{1}$$

A physician's utility increases with profits π and an altruistic part that accounts for the patient benefit, i.e. $\theta B(q)$. Utility is therefore

$$V(\theta, q) = \pi + \theta B(q) . \tag{2}$$

Moreover, physicians care about their reputation among patients and other physicians. Similar to Siciliani (2009) we assume that a physician is perceived as a "good" physician if the provided medical treatment q is within a reputation interval $[\underline{q}, \overline{q}]$ with $q_B^* \in [\underline{q}, \overline{q}]$. Sufficiently strong underprovision $q < \underline{q}$ or overprovision $q > \overline{q}$ yield zero reputation. As

[2] Note that q is not a measure for quality or performance as assumed in Eggleston (2005) or Siciliani (2009). Quite in contrary, we focus on overprovision where a higher q implies lower quality for the patient.

[3] Alternatively, it may also be interpreted as intrinsic motivation, see Besley and Gathak (2005).

physicians enjoy being regarded as good they receive an extra gain in utility from reputation which we model as

$$(\alpha + \lambda\delta p)\, w > 0 \tag{3}$$

with $\alpha, \delta, w > 0$ and $\lambda = 1$ if $q \geq q_B^*$ and $\lambda = -1$ otherwise.[4]

The higher α the more a physician enjoys being perceived as good. A higher δ reflects a stronger stigma associated with providing care under financial incentives. If a physician provides $q^* < q_B^*$ such that $\lambda = -1$, a higher p devalues reputation. In this case a higher q not only reflects providing better medical treatment but also has a negative stigma associated with financial incentives (greediness) to maximize revenues, see Le Grand (2003). The situation is the opposite in case of overprovision with $q^* > q_B^*$ and $\lambda = 1$. In this case a physician providing less medical treatment can unambiguously be identified as a physician that cares more about patients and less about maximizing revenues.

The specification of reputation is discrete. This assumption may be justified as patients' judgments about physicians' abilities are not extremely accurate and they usually call them a "good" or "bad" physician. Like Siciliani (2009) we assume the dichotomous case, i.e., physicians are either good or not good. If a physician provides $q \in [\underline{q}, \overline{q}]$ the extra gain from reputation yields a utility level of

$$U\left(\theta, q \in [\underline{q}, \overline{q}]\right) = V(\theta, q) + (\alpha + \lambda\delta p)\, w. \tag{4}$$

In contrast, if the physician is not regarded as good utility $U(\theta, q \notin [\underline{q}, \overline{q}])$ is given by (2).

Now, consider a physician's maximization problem. The optimal medical treatment $q^*(\theta)$ in case there is no extra utility gain from reputation $(w = 0)$ is implicitly given by

$$p + \theta B_q = C_q \tag{5}$$

with

$$\frac{\partial q^*}{\partial \theta} = \frac{B_q}{C_{qq} - \theta B_{qq}}. \tag{6}$$

More altruistic physicians provide relatively more (less) in case of underprovision (overprovision) since $q^* < q_B^* \Rightarrow B_q > 0$ ($q^* > q_B^* \Rightarrow B_q < 0$). Moreover, it follows

$$\frac{\partial q^*}{\partial p} = \frac{1}{C_{qq} - \theta B_{qq}} > 0 \tag{7}$$

[4] As in Siciliani (2009) we assume $\delta p < \alpha$ for $\lambda = -1$ to secure that the reputation gain is positive.

such that a higher FFS increases the likelihood of overprovision. In the following we focus on overprovision and assume that p is sufficiently high such that $q^* > q_B^*$ for all θ. Note that if p is sufficiently low such that $q^* < q_B^*$ for all θ we consider the case of Siciliani (2009).

We now explore how altruism shapes the physician's utility. In Figure 1 we focus on three types of physicians who differ in their degree of altruism $\theta_3 > \theta_2 > \theta_1$, with θ_3 being the type with the highest degree of altruism. Due to the extra gain from reputation a physician's utility function $U(\theta, q)$ has three discontinuities. The first one at $q = \underline{q}$ and the one third at $q = \overline{q}$. The utility jumps upwards (downwards) when a quantity weakly above $\underline{q}$ ($\overline{q}$) is provided. The jump is due to the extra utility gain and equal to $(\alpha + \lambda\delta p)\, w$. Moreover, the utility jumps up at $q = q_B^*$ since the sign of λ changes.

In the following we show that all types of physicians can be grouped into three categories: i.) *high*, ii.) *intermediate* and iii.) *low* altruism. High altruism physicians provide quantity $q^*(\theta) \in [\underline{q}, \overline{q}]$ - independent of whether reputation yields an extra utility gain or not. This case is illustrated by type θ_3 in Figure 1. If the physician chooses quantity $q^*(\theta_3)$, she obtains a utility (point C) which is higher then the utility she would obtain if quantity $\tilde{q} \equiv \overline{q}$ was chosen (point C'). Define $\tilde{\theta}$ as the level of altruism such that the provider is indifferent between $q^*(\tilde{\theta})$ and $\tilde{q}$. We assume that $q^*(\overline{\theta}) < \tilde{q}$, otherwise the group with high altruism would be empty. Then, physicians of high altruism type in the range $\tilde{\theta} < \theta < \overline{\theta}$ provide output $q^*(\theta)$ and receive utility $V(\theta, q^*(\theta)) + (\alpha + \delta p)\, w$.

If the physician's degree of altruism is below the threshold $\tilde{\theta}$, it follows $q^*(\theta) > \tilde{q}$. Now the physician faces a trade-off. If she provides $\tilde{q}$, she gains a good reputation which increases her utility by $(\alpha + \delta p)\, w$. However, providing $\tilde{q}$ is costly (in terms of foregone revenue) as it is below $q^*(\theta)$. The physician provides $\tilde{q}$ if

$$(\alpha + \delta p)\, w > V(\theta, q^*(\theta)) - V(\theta, \tilde{q}), \tag{8}$$

i.e., if the additional utility from being perceived as a good physician is higher than the loss in utility from choosing quantity $\tilde{q}$ instead of $q^*(\theta)$. Since

$$\frac{\partial\left[V(\theta, q^*(\theta)) - V(\theta, \tilde{q})\right]}{\partial\theta} = -\left[B(\tilde{q}) - B(q^*(\theta))\right] < 0 \tag{9}$$

we can conclude that physicians with a higher degree of altruism have a lower loss of utility from choosing $\tilde{q}$. We assume that for the physician with the lowest degree of altruism it is not optimal to provide quantity $\tilde{q}$. Then, there exists a level of altruism $\hat{\theta}$ defined by

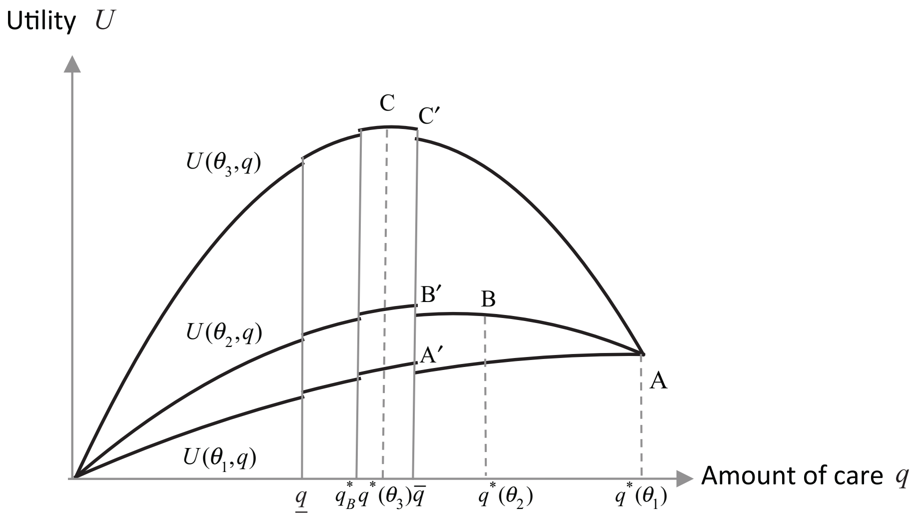


Figure 1: Physician's utility function for different degrees of altruism

$V(\hat{\theta}, q^*(\theta)) - V(\hat{\theta}, \tilde{q}) = (\alpha + \delta p)\, w$ such that physicians with a degree of altruism below (above) $\hat{\theta}$ choose quantity $q^*(\theta)$ $(\tilde{q})$. We refer to physicians in the first group as low altruism physicians, and to the second group as physicians with intermediate levels of altruism.

As illustrated in Figure 1, physician θ_1 obtains a higher utility by choosing quantity $q^*(\theta_1)$ (point A) rather than quantity $\tilde{q}$ (point A'). Physician θ_1 therefore belongs to the category of low-altruism physicians. In contrast, physician θ_2 obtains a higher utility by choosing quantity $\tilde{q}$ (point B') rather than quantity $q^*(\theta_2)$ (point B). Physician θ_2 belongs to the category of physicians with intermediate altruism. Notice that even if a physician has a higher (but still intermediate) degree of altruism compared to provider θ_2, she provides the same quantity $\tilde{q}$. Figure 2 illustrates the three different altruism groups and their optimal treatment levels for over- and underprovision. All groups overprovide (underprovide) while the magnitude of overprovision (underprovision) is most severe for the low altruism group.

As in Siciliani (2009) we consider the total amount of care across physicians

$$Q(p, w) = \int_{\underline{\theta}}^{\hat{\theta}(p,w)} q^*(\theta, p)\, f(\theta)\, \mathrm{d}\theta + \int_{\hat{\theta}(p,w)}^{\tilde{\theta}(p,w)} \tilde{q} f(\theta)\, \mathrm{d}\theta + \int_{\tilde{\theta}(p,w)}^{\overline{\theta}} q^*(\theta, p)\, f(\theta)\, \mathrm{d}\theta \qquad (10)$$

and derive comparative statics with respect to reputational motivation and the FFS. First,

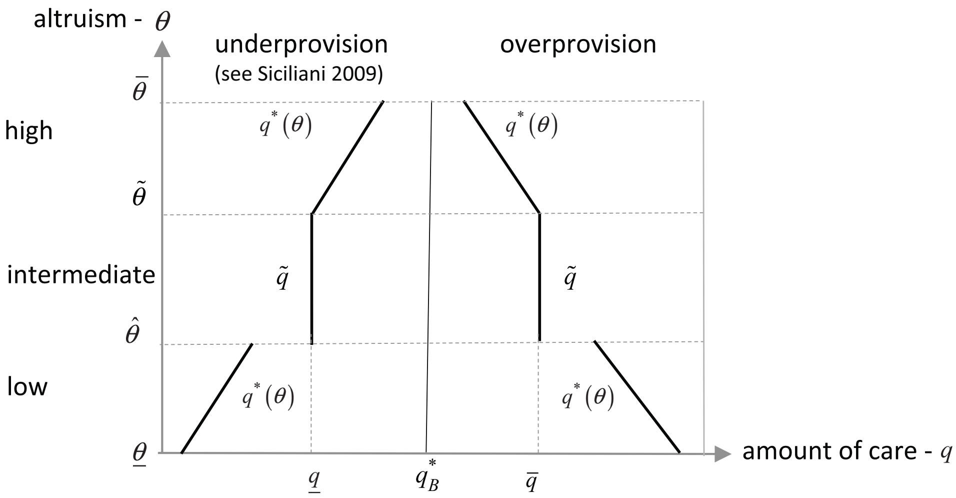


Figure 2: Amount of care for different types.

assuming an uniform distribution for the degrees of altruism we obtain[5]

$$\frac{\mathrm{d}Q}{\mathrm{d}w} = -\frac{(\alpha + \delta p)\,(q^*(\hat{\theta}) - \tilde{q})}{B\,(\tilde{q}) - B(q^*(\hat{\theta}))} < 0. \tag{11}$$

Intuitively, better reputational motivation increases the incentive for low-altruism physicians to provide $\tilde{q}$ which in turn decreases $\hat{\theta}$. As illustrated in Figure 3 this reduces the magnitude of overprovision (area C).[6] This implies the important result that the (variable) costs of healtcare provision ($p\,Q$) decrease with better reputational motivation.[7]

Second, a higher FFS has an ambiguous effect on Q

$$\frac{\mathrm{d}Q}{\mathrm{d}p} = \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial q^*(\theta, p)}{\partial p}\mathrm{d}\theta + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial q^*(\theta, p)}{\partial p}\mathrm{d}\theta + \frac{q^*(\hat{\theta}) - \tilde{q}}{B(\tilde{q}) - B(q^*(\hat{\theta}))}\left[q^*(\hat{\theta}) - \tilde{q} - \delta w\right]. \tag{12}$$

In case the extra gain of reputation is sufficiently large, i.e., if $\delta w > q^*(\hat{\theta}) - \tilde{q}$, an increase in p can lead to a lower Q. Intuitively, a higher p has two effects. First, an increase in p induces physicians with low and high altruism to increase output. Second, it changes

[5]See Appendix i) for details.

[6]Moreover, we show in Appendix ii) that in case of underprovision (see Siciliani (2009)), better motivational reputation increases the quantity of medical treatment and reduces the magnitude of underprovision.

[7]It is clear that better reputational motivation may involve extra administrative costs which may dampen the reduction.

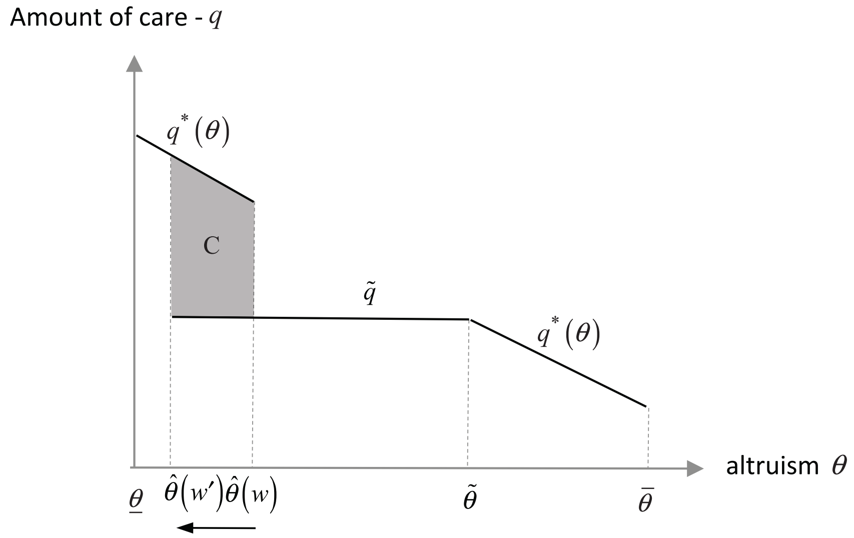


Figure 3: Effect of an increase in motivational reputation on the amount of care.

$\hat{\theta}$ and $\tilde{\theta}$.[8] Both intermediate and high altruism physicians receive the extra reputation gain, while the latter group increases q^*. Hence, the cutoff $\tilde{\theta}$ increases and less physicians are considered as high altruism type. The change in physicians between the low- and intermediate-altruism group is ambiguous. On the one hand, a higher p makes it less attractive for intermediate-altruism physicians to provide quantity $\tilde{q}$ due to the foregone revenue. On the other hand, an increase in p also increases the reputation gain. If the latter effect dominates, an increase in p reduces Q. Graphically, the reduction in quantity (area C in Figure 4) offsets the increase in quantity (area $A + B$). Nevertheless, it is important to note that a lower p may actually increase the costs of healthcare provision.

3 Policy regimes

The standard policy regime is to set a FFS. As a complementary instrument, policy makers may also induce better reputational motivation to decrease costs. In the following we derive p^* that maximizes the "efficiency" of the incentive scheme (a weighted measure W of physicians' utility U^* and patients' benefit B^*) and show that p^* decreases with better

[8]Note that a change in $\tilde{\theta}$ or $\hat{\theta}$ does not imply that a physician's exogenously given degree of altruism changes. However, an increase in p changes the endogenous classifications, e.g., a lower $\hat{\theta}$ implies that some former low altruism physicians are considered to be of intermediate altruism type.

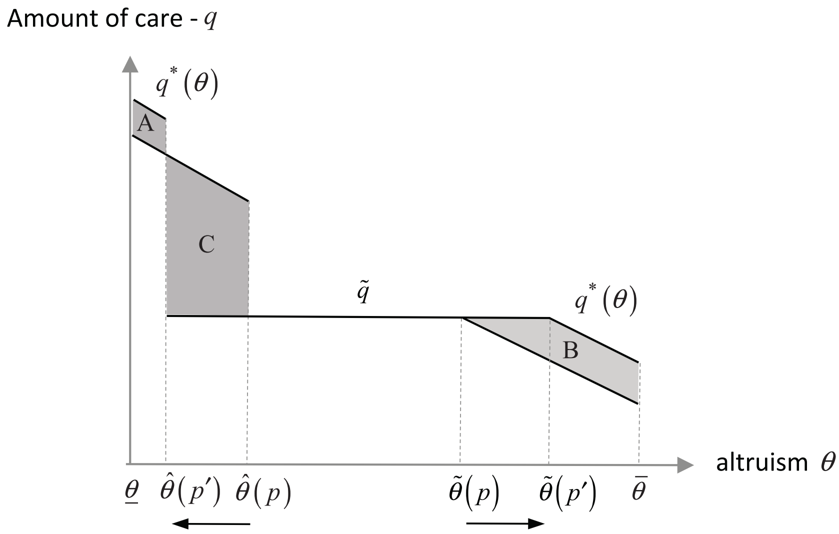


Figure 4: Effect of an increase in the FFS on the amount of care.

reputational motivation. The efficiency of incentive scheme is given by

$$W = \sigma\, U^{*} + (1-\sigma)\, B^{*} \tag{13}$$

with $\sigma \in (0,1)$ and where a higher (lower) σ puts a stronger weight on physicians' utility $U^{*} \equiv \int_{\underline{\theta}}^{\overline{\theta}} U(\theta)\,\mathrm{d}\theta$ (patients' benefit $B^{*} \equiv \int_{\underline{\theta}}^{\overline{\theta}} B(\theta)\,\mathrm{d}\theta$).[9]

Physicians aim for a higher FFS (see Appendix iii) for details) since

$$\frac{\mathrm{d}U^{*}}{\mathrm{d}p} = Q + \int_{\hat{\theta}}^{\overline{\theta}} \delta w \mathrm{d}\theta > 0. \tag{14}$$

First, a higher p increases the marginal revenue of care (Q). Second, the reputation effect $\partial(\alpha+\delta p)w/\partial p = \delta w$ is positive since being regarded as a "good" physician yields a greater extra utility gain. Conversely, patients prefer a lower FFS since it reduces the magnitude of overprovision. Formally,

$$\frac{\mathrm{d}B^{*}}{\mathrm{d}p} = \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial B}{\partial q^{*}}\frac{\partial q^{*}}{\partial p}\mathrm{d}\theta + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial B}{\partial q^{*}}\frac{\partial q^{*}}{\partial p}\mathrm{d}\theta - \left[q^{*}(\hat{\theta}) - \tilde{q} - \delta w\right] < 0 \tag{15}$$

[9]As Siciliani (2009) we do not consider a budget constraint since we focus on the trade-off between physicians' utilities and patients' benefits. Nevertheless, a budget constraint would imply that p^{*} would be bounded from above.

where the two integrals are negative since a higher FFS increases the quantity of low- and high-altruism providers.[10]

Using (14) and (15) the efficiency maximizing p^* (see Appendix iv) for details) satisfies

$$\frac{\mathrm{d}W}{\mathrm{d}p} = \sigma \, \frac{\partial U^*}{\partial p} + (1-\sigma) \, \frac{\partial B^*}{\partial p} = 0. \tag{16}$$

Using (16) allows us to derive comparative statics for p^* with respect to w. We show (see Appendix iv) for details) that $\partial p^*/\partial w = -\left(\partial W_p/\partial w\right)\left(\partial W_p/\partial p\right)^{-1} < 0$. In words, an increase in reputational motivation decreases the efficiency maximizing FFS.[11]

4 Conclusion

We have provided a model where physicians differ in their degree of altruism and that explicitly allows for overprovision of medical treatment. Our main result is that in case of overprovision better reputational motivation decreases overall output and the efficiency maximizing FFS. Promoting reputational motivation may therefore both decrease the costs in the healthcare system and simultaneously increase patient benefits.

[10] We assume that the group of low- and high-altruism physicians is sufficiently large such that the first two terms dominate the third term which can become negative if δw is sufficiently large. Otherwise an increase in p has a negligible effect on the total amount of care. In the extreme case of $p \to \infty$ there are no low- and high-altruism providers which we rule out since the comparative statics are then meaningless.

[11] In Appendix v) we show that in case of underprovision both physicians and patients aim for a higher FFS. This implies that policy makers are likely to find themselves in an overprovision scenario.

References

Bénabou, R. and J. Tirole (2006): Incentives and Prosocial Behavior, *American Economic Review*, 96, 1652-1678.

Besley, T. and M. Gathak (2005): Competition and Incentives with Motivated Agents, *American Economic Review*, 95(3), 616–636.

Chalkley, M. and F. Khalil (2005): Third Party Purchasing of Health Services: Patient Choice and Agency, *Journal of Health Economics*, 24(6), 1132-1153.

Cutler, D.M., and D.P. Ly (2011): The (Paper)Work of Medicine: Understanding International Medical Costs, *Journal of Economic Perspectives*, 25(2), 3–25.

Doran, T., Fullwood,C., Gravelle, H., Reeves, D., Kontopantelis, E., Hiroeh,U. and M. Roland (2006): Pay-for-Performance Programs in Family Practices in the United Kingdom, *New England Journal of Medicine*, 355, 375-384.

Dranove, D., Kessler, D., McClellan, M. and M. Satterthwaite (2003): Is More Information Better? The Effects of "Report Cards" on Health Care Providers, *Journal of Political Economy*, 111(3), 555-588.

Eggleston, K. (2005): Multitasking and Mixed Payment Systems for Provider Payment, *Journal of Health Economics*, 24, 211-223.

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

Glazer, J., McGuire, T.G. and J.P. Newhouse (2007): Using Performance Measures to Motivate 'Report-averse' and 'Report-loving' Agents, *Journal of Health Economics*, 26(6), 1170-1189.

Le Grand, J. (2003): Motivation, Agency and Public Policy: of Knights and Knaves, Pawns and Queens, *Oxford University Press*, Oxford.

McClellan, M. (2011): Reforming Payments to Healthcare Providers: The Key to Slowing Healthcare Cost Growth while Improving Quality?, *Journal of Economic Perspectives*, 25(2), 69-92.

McGuire, T. G. (2000): Physician Agency, in *Handbook of Health Economics*, Vol. 1 A, ed. by A. C. Cuyler and J. P. Newhouse, Amsterdam, North-Holland.

Rosenthal, M.B., Fernandopulle, R., Song, H.R., and B. Landon (2004): Paying for Quality: Providers' Incentives for Quality Improvement, *Health Affairs*, 23(2), 127-41.

Siciliani, L. (2009): Paying for Performance and Motivation Crowding Out, *Economics Letters*, 103, 68-71.

Werner, R.M., and D. A. Asch (2005): The Unintended Consequences of Publicly Reporting Quality Information, *Journal of the American Medical Association*, 293(10), 1239-1244.

Appendix

i) *Total amount of care* is given by Q as stated in (10). To see that the change with respect to p is given by (12) consider

$$\begin{aligned}\frac{\mathrm{d}Q}{\mathrm{d}p} &= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta + q^*(\hat{\theta})\frac{\partial\hat{\theta}}{\partial p} + \int_{\hat{\theta}}^{\tilde{\theta}} \frac{\partial \tilde{q}}{\partial p}\mathrm{d}\theta + \tilde{q}\frac{\partial\tilde{\theta}}{\partial p} - \tilde{q}\frac{\partial\hat{\theta}}{\partial p} + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta - q^*(\tilde{\theta})\frac{\partial\tilde{\theta}}{\partial p} \\ &= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta + + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta + \frac{\partial\hat{\theta}}{\partial p}\left[q^*(\hat{\theta}) - \tilde{q}\right] \qquad (17)\end{aligned}$$

and note that the indifference condition given by (8) implies

$$\frac{\partial\hat{\theta}}{\partial p} = \frac{q^*(\hat{\theta}) - \tilde{q} - \delta w}{B(\tilde{q}) - B(q^*(\hat{\theta}))} < 0 \qquad (18)$$

if δw is sufficiently large, see Siciliani (2009) for this assumption. Using (18) in (17) yields (12).

ii) *Comparison to Siciliani (2009)*: We claim that in case of underprovision better reputational motivation increases the overall quantity of medical treatment and reduces the magnitude of underprovision. To consider underprovision, we assume for the moment that the FFS p is sufficiently low such that $q < q_B^*$ for all degrees of altruism θ. Now consider a higher w. The change in Q is given by

$$\frac{\mathrm{d}Q}{\mathrm{d}w} = \frac{\partial\hat{\theta}}{\partial w}\left[\tilde{q} - q^*(\hat{\theta})\right] = \frac{(\alpha - \delta p)}{B(\tilde{q}) - B(q^*(\hat{\theta}))}\left(\tilde{q} - q^*(\hat{\theta})\right) > 0. \qquad (19)$$

It is unambiguously positive since in case of underprovision we have i.) $B(\tilde{q}) > B(q^*(\hat{\theta}))$, ii.) $\tilde{q} > q^*(\hat{\theta})$ and iii.) $\alpha > \delta p$, see Siciliani (2009) for a detailed discussion of those properties. Intuitively, a better general reputational motivation unambiguously increases the incentive to provide more medical treatment. In contrast to the case of overprovision no physician has an incentive to decrease the amount of medical treatment.

iii) *Efficiency maximizing FFS* p^*: First, aggregated physicians utility is given by

$$U^* \equiv \int_{\underline{\theta}}^{\overline{\theta}} U(\theta)\,\mathrm{d}\theta. \tag{20}$$

Physicians unambiguously aim for a higher price since

$$\begin{aligned}
\frac{\mathrm{d}U^*}{\mathrm{d}p} &= \int_{\underline{\theta}}^{\hat{\theta}} q^*\mathrm{d}\theta + V(q^*(\hat{\theta}))\frac{\partial\hat{\theta}}{\partial p} + \int_{\hat{\theta}}^{\tilde{\theta}} \tilde{q}\mathrm{d}\theta - V(\tilde{q})\frac{\partial\hat{\theta}}{\partial p} + V(\tilde{q})\frac{\partial\tilde{\theta}}{\partial p} \\
&\quad + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta - V(q^*(\tilde{\theta}))\frac{\partial\tilde{\theta}}{\partial p} + \int_{\hat{\theta}}^{\overline{\theta}} \delta w\mathrm{d}\theta - (\alpha+\delta p)\,w\frac{\partial\hat{\theta}}{\partial p} \\
&= Q + \int_{\hat{\theta}}^{\overline{\theta}} \delta w\mathrm{d}\theta + \frac{\partial\hat{\theta}}{\partial p}\left[V(q^*(\hat{\theta})) - V(\tilde{q}) - (\alpha+\delta p)\,w\right] \\
&= Q + \int_{\hat{\theta}}^{\overline{\theta}} \delta w\mathrm{d}\theta > 0.
\end{aligned} \tag{21}$$

Note that the last term in the brackets is zero due to the indifference condition given by (8). Second, aggregated patient benefit is given by

$$B^* \equiv \int_{\underline{\theta}}^{\overline{\theta}} B(\theta)\,\mathrm{d}\theta \tag{22}$$

and patients aim for a lower FFS since

$$\begin{aligned}
\frac{\mathrm{d}B^*}{\mathrm{d}p} &= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta + B(q^*(\hat{\theta}))\frac{\partial\hat{\theta}}{\partial p} - B(\tilde{q})\frac{\partial\hat{\theta}}{\partial p} + B(\tilde{q})\frac{\partial\tilde{\theta}}{\partial p} + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta - B(q^*(\tilde{\theta}))\frac{\partial\tilde{\theta}}{\partial p} \\
&= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta + + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta + \frac{\partial\hat{\theta}}{\partial p}\left[B(q^*(\hat{\theta})) - B(\tilde{q})\right] < 0 \\
&= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta - q^*(\hat{\theta}) + \tilde{q} + \delta w < 0
\end{aligned} \tag{23}$$

if the group of high- and low-altruism physicians is large enough. As explained in footnote 13 this is our general assumption to secure that the comparative static results are meaningful. Third, consider the second-order-conditions. In order to secure that p^* which solves (16) actually is a maximum, we have to show that $W_{pp} = \sigma B_{pp} + (1-\sigma)U_{pp} < 0$ for $p = p^*$. Hence, consider

$$\begin{aligned}
\frac{\mathrm{d}U_p^*}{\mathrm{d}p} &= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta + \frac{q^*(\hat{\theta}) - \tilde{q}}{B(\tilde{q}) - B(q^*(\hat{\theta}))}\left[q^*(\hat{\theta}) - \tilde{q} - \delta w\right] - \delta w\left[\frac{q^*(\hat{\theta}) - \tilde{q} - \delta w}{B(\tilde{q}) - B(q^*(\hat{\theta}))}\right] \\
&= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta + \frac{(q^*(\hat{\theta}) - \tilde{q} - \delta w)^2}{B(\tilde{q}) - B(q^*(\hat{\theta}))} > 0
\end{aligned} \tag{24}$$

which is unambiguously positive. Hence, if σ is low such that U_{pp} yields $W_{pp} > 0$, the solution of $W_P = 0$ actually is a minimum. In this case physicians' will for a higher price dominates the decrease in the patients' benefit. Now consider the second-order-condition of patient benefit which is given by

$$\frac{\mathrm{d}B_p^*}{\mathrm{d}p} = \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial B_p}{\partial p}\mathrm{d}\theta + B_p(\hat{\theta})\frac{\partial\hat{\theta}}{\partial p} + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial B_p}{\partial p}\mathrm{d}\theta - B_p(\tilde{\theta})\frac{\partial\tilde{\theta}}{\partial p} - \frac{\partial q^*(\hat{\theta})}{\partial\hat{\theta}}\frac{\partial\hat{\theta}}{\partial p} < 0 \tag{25}$$

where $B_p < 0$, $B_{pp} < 0$, $\partial\tilde{\theta}/\partial p > 0$, $\partial q^*(\hat{\theta})/\partial\hat{\theta} < 0$, and $\partial\hat{\theta}/\partial p < 0$ is given by (18). Again we assume that the group of high- and low-altruism physicians is large enough such that B_{pp}^* is unambiguously negative. This is important to secure that for a sufficiently high σ a p^* that solves (16) actually constitutes a maximum since then we have $W_{pp} < 0$. Note that if either the group of high- and low-altruism physicians is small or σ is low a higher FFS does not necessarily constitute a trade-off between patients and physicians. If σ is low (physicians' utilities are most important) the efficiency maximizing p^* would be infinitesimally high while if σ is high (patients' benefits are most important) p^* would be infinitesimally small.

iv) *Reputational motivation and* p^*: In the text we claim that better reputational motivation decreases the efficiency maximizing FFS p^*. To show this result consider

$$\frac{\partial B_p}{\partial w} = B_p\frac{\partial\hat{\theta}}{\partial w} - \frac{q^*(\hat{\theta})}{\partial\hat{\theta}}\frac{\partial\hat{\theta}}{\partial w} + \delta < 0 \tag{26}$$

which is negative if the change in $\partial q^*/\partial\hat{\theta} < 0$ is sufficiently strong. Furthermore, consider

$$\frac{\partial U_p^*}{\partial w} = \frac{\mathrm{d}Q}{\mathrm{d}w} + \int_{\hat{\theta}}^{\overline{\theta}} \delta\mathrm{d}\theta - \delta w\frac{\partial\hat{\theta}}{\partial w} < 0 \tag{27}$$

where $\mathrm{d}Q/\mathrm{d}w$ is given by (11) and again it is assumed that the group of high- and low-altruism physicians is large enough. This secures that $\partial U_p^*/\partial w <$ is unambiguously negative. Taken together we have $\partial W_p/\partial w < 0$ and $W_{pp} < 0$ such that $\partial p^*/\partial w = -(\partial W_p/\partial w)/W_{pp} < 0$. An analogous approach yields the comparative static results for α and δ.

v) *Comparison to Siciliani (2009)*: Similar to the case of overprovision, in case of underprovision physicians also aim for a higher price since

$$\begin{aligned}
\frac{\mathrm{d}U^*}{\mathrm{d}p} &= \int_{\underline{\theta}}^{\hat{\theta}} q^*\mathrm{d}\theta + V(q^*(\hat{\theta}))\frac{\partial\hat{\theta}}{\partial p} + \int_{\hat{\theta}}^{\tilde{\theta}} \tilde{q}\mathrm{d}\theta - V(\tilde{q})\frac{\partial\hat{\theta}}{\partial p} + V(\tilde{q})\frac{\partial\tilde{\theta}}{\partial p} \\
&\quad + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial q^*}{\partial p}\mathrm{d}\theta - V(q^*(\tilde{\theta}))\frac{\partial\tilde{\theta}}{\partial p} - \int_{\hat{\theta}}^{\overline{\theta}} \delta w\mathrm{d}\theta - (\alpha - \delta p)\, w\frac{\partial\hat{\theta}}{\partial p} \\
&= Q - \int_{\hat{\theta}}^{\overline{\theta}} \delta w\mathrm{d}\theta + \frac{\partial\hat{\theta}}{\partial p}\left[V(q^*(\hat{\theta})) - V(\tilde{q}) - (\alpha - \delta p)\, w\right] \\
&= Q - \int_{\hat{\theta}}^{\overline{\theta}} \delta w\mathrm{d}\theta > 0
\end{aligned} \tag{28}$$

if the greediness "penalty" δw is not too strong. Next, consider the patients' benefits given by

$$\begin{aligned} \frac{\mathrm{d}B^*}{\mathrm{d}p} &= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta + \frac{\partial \hat{\theta}}{\partial p}\left[B(q^*(\hat{\theta})) - B(\tilde{q})\right] \\ &= \int_{\underline{\theta}}^{\hat{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta + \int_{\tilde{\theta}}^{\overline{\theta}} \frac{\partial B}{\partial q^*}\frac{\partial q^*}{\partial p}\mathrm{d}\theta + \delta w + \tilde{q} - q^*(\hat{\theta}) > 0 \end{aligned} \tag{29}$$

which is unambiguously positive since in the case of underprovision a higher FFS mitigates the underinvestment problem, i.e., $B_p > 0$ and not $B_p < 0$ as in the case of overprovision. Hence, both physicians and patients aim for a higher FFS in case of underprovision which can unambiguously be seen by (28) and (29), respectively. This in turn implies that in the case of underprovision an efficiency maximizing FFS p^* which solves $W_p = 0$ as given by (16) cannot exist.

Conclusion

Conclusion

It is well documented in the literature that incentives in the workplace affect agents' behavior, though the way they do varies greatly across production environments (Pendergast, 1999). Designing the "right lever", as Levitt and Dubner (2009) put it, is thus of particular importance. Studying incentives empirically is difficult, however, and evidence from more complex production situations often lacks controlled analysis so that changes in behavior cannot necessarily be attributed to the incentives. The five studies discussed here consider two types of complex situations: complex team production processes in which marginal productivities of the individuals are not always clear, and the complex decision environment faced by physicians, where performance in terms of outcome quality is usually difficult to observe. These studies use laboratory experiments and corresponding theoretical models to shed light on the underlying mechanisms of incentives in these complex situations, while allowing for a controlled analysis. The findings of these studies can be summarized in three main results.

Result 1

The degree of obviousness of performance measures in complex team production situations plays an important role for team members' perceptions of distributive fairness.

In Chapter 1 we analyze perceptions of distributive fairness in complex joint production tasks. Joint production is complex in the sense that individual contributions to the joint output are not always apparent, as a result of potential negative externalities or synergies of the production. We find that if marginal productivities are noticeable, but not obvious, team members do not take them into account. The most prevalent distribution norm in this case is the accountability principle. Only when individual marginal productivities are made highly apparent do people take them into account. Our results thus imply that in complex team production designers of payment incentives should be sensitive to the degree of salience of the performance measures. In cases where individual productivities are not obvious, performance incentives should account for observable inputs only.

RESULT 2

Subjects in the role of physicians react significantly to financial and non-financial incentives and improve patient outcomes.

In Chapters 2-4 we analyze monetary and non-monetary incentives for physicians. To this end, subjects in the experiments take on the role of physicians and decide on the quantity of medical treatment, knowing that the monetary patient benefit will be transferred to real patients outside the laboratory. First we analyze physician provision behavior under monetary incentives, given a payment reform from basic capitation or fee-for-service remuneration to either mixtures of both or additional P4P incentives. We find that both types of new payment incentives affect provision behavior, leading to treatment levels significantly closer to the patient optimum. Hence, if policy makers wish to improve the quality of health care in terms of better treatment of patients, mixed and P4P incentives can be a successful means. However, we also find that efficiency – in terms of patient benefit per unit of money spent – does not improve under mixed remuneration with a constant profit maximum and under P4P incentives with an additional bonus. Only lowering the profit maximum of the mixed remuneration scheme also leads to an increase in the benefit-cost ratio. Thus, if policy makers seek not only to improve patient benefits but also to increase efficiency, they must lower the maximum profit level given mixed remuneration. Second, we analyze physician provision behavior in a reform from basic fee-for-service remuneration to additional non-monetary incentives. The latter include private feedback, which speaks to a person's self-esteem, and public feedback, addressing reputational motivation. We find that while public feedback significantly improves the quality of care provided, private feedback has no effect on the latter. Concerning the design of future non-monetary performance incentives, this suggests that only public feedback incentives may lead to an increase in treatment quality.

RESULT 3

In a theoretical framework in which financial and non-financial incentives are in place simultaneously, only an increase in the latter unambiguously affects physician provision behavior such that health care costs are reduced and patient benefit is increased.

In a theoretical model I consider the case in which financial and non-financial incentives are jointly implemented, as is the case in most P4P programs. In the theoretical model physicians differ in their degree of altruism and in the extent to which they care about their reputation. In particular, I focus on the case of overprovision of medical treatment, as improving patient benefit and lowering costs should not be mutually exclusive aims here. The theoretical results indicate that, in contrast to monetary incentives, improving non-monetary reputation mechanisms unambiguously reduces the costly overprovision of medical treatment and improves patient benefits. Moreover, when introduction costs are neglected, increasing reputation incentives decreases the welfare maximizing price in a fee-for-service scheme. For policy makers reputation-based incentive mechanisms may thus be an efficient means to simultaneously improving patient care and reducing healthcare costs.

Results 1-3 underline the importance of controlled laboratory experiments in complex production situations in the workplace. The five studies of this dissertation contribute substantially to the empirical evidence in this area by yielding controlled analyses of performance and payment incentives in complex production situations where field data is difficult to collect. They also indicate interesting perspectives for future research.

FUTURE RESEARCH

Chapter 1 provides pertinent results for remuneration in multiple business practices, from joint ventures to co-authored research projects. Future research could target the limitations of the study’s design. In particular, since we use multiple choice questions as real effort task, individual productivities are not observable and we cannot differentiate between effort, talent, and luck. Consequently, we are unable to distinguish the accountability principle from liberal egalitarianism. Future experiments could therefore use a design which controls for talent and / or intention, for example by letting subjects determine the level of difficulty of the questions they have to answer. Moreover, the perceptions of distributive fairness might be different in the domain of losses and in the domain of gains.

The results presented in chapters 2-4 on the underlying mechanisms of physician payment and performance incentives are useful for health care policy makers. The experimental design developed and used in Chapters 2-4 is highly innovative and may serve as a basis for future research on payment mechanisms and performance incentives for physicians. One suggestion for future experiments is to gradually increase the complexity of the production process in this context. A key characteristic of a physician's decision environment is the uncertainty about the outcome of the treatment (Arrow, 1963). This may affect not only the patient's benefit, but also the physicians' profits if they are based on patient outcomes (as is the case in P4P schemes). In Chapters 2-4, we assumed both patient benefit and physician profit to be fully observable. In order to analyze whether P4P incentives also lead to better patient outcomes in situations under risk, one could extend the basic design to include uncertainty over the patient outcome and over the physician's profit. In such a scenario with uncertainty over the outcome, physicians might treat patients differently and P4P incentives may be less effective.[1] A second angle for future experiments is to allow for multitasking, i.e. by letting physicians make both incentivized and non-incentivized treatment decisions for different patients under a capacity constraint. Multitasking is a critical problem of performance incentives, and physicians might react by increasing treatment quality for patients for whom they receive an incentive, while lowering treatment quality for patients where they do not face performance rewards (Kaarbøe and Siciliani, 2010). In Chapter 3, we assumed that quality is measureable for all patients. When integrating capacity constraints into our design, one could analyze whether physicians substitute incentivized patients for non-incentivized ones, and whether P4P incentives then still lead to an overall increase in quality.

Chapter 5 provides novel testable predictions for future laboratory experiments. In particular, one could analyze the effect of jointly introducing the P4P incentives of Chapter 2 and the non-monetary public feedback of Chapter 4. This is especially interesting as these two policy means are usually implemented together, e.g. in the Quality and Outcomes Framework in the UK. As the effects of these two types of incentives are often difficult to disentangle, it is unclear whether they are substitutes or complements (Maynard, 2012), or whether monetary incentives even crowd out

[1] For first results see Brosig-Koch et al. (2013, mimeo dggö).

reputational motivation mechanisms – as suggested by the theoretical results of Chapter 5 and by the experimental results of Ariely et al. (2009) in a non-medical setting.

Once the complexity of a physician's profession is analyzed more comprehensively, future experimental designs may provide a useful and efficient future test bed for policy makers. By testing payment reforms in the laboratory rather than in the field, policy makers can avoid costly failures and possible disadvantages for certain patient types.

Literature

Ariely, D., Bracha, A., and S. Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99, 544–55.

Arrow, K.J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53, 941-973.

Brosig-Koch, J., H. Hennig-Schmidt, and N. Kairies (2013): "Medical treatment with unknown patient outcome: an experiment." mimeo dggö 2013.

Kaarbøe, O. and L. Siciliani, 2010. "Multitasking, quality and pay for performance." *Health Economics* 20, 225–238.

Levitt, S. D. and S. J. Dubner. 2009. "Superfreakonomics." William Morrow, New York.

Prendergast, C. J. 1999. "The provision of incentives in firms." *Journal of Economic Literature* 37, 7-63.

Maynard, A. 2012. "The powers and pitfalls of payment for performance." *Health Economics* 21(1), 3-12.

Attachments

Überblick über die Einzelbeiträge

- Brosig-Koch, J., Hennig-Schmidt, H., Kairies, N, and D. Wiesen. 2013a. "How to Improve Patient Care? An Analysis of Capitation, Fee-for-Service, and Mixed Incentive Schemes for Physicians."
  Diese Arbeit ist 2013 als *Ruhr Economic Paper* No. 412 erschienen.

- Brosig-Koch, J., Hennig-Schmidt, H., Kairies, N., and D. Wiesen. 2013b. "How Effective are Pay-for-Performance Incentives for Physicians? A Laboratory Experiment."
  Diese Arbeit ist 2013 als *Ruhr Economic Paper* No. 413 erschienen.

- Fischbacher, U., Kairies, N., and U. Stefani. 2013. "Equal Sharing, Accountability, and Productivity: An Experiment on Distributive Fairness in Complex Team Production Processes."
  Diese Arbeit ist beim *Journal Economic Behavior & Organization* eingereicht.

- Kairies, N. 2012. "P4P, Reputation and the Reduction of Costly Overprovision."
  Diese Arbeit ist in *Economics Bulletin*, Vol. 33 No.3, p. A211 erschienen.

- Kairies, N. and M. Krieger. 2013. "How do Non-Monetary Performance Incentives for Physicians Affect the Quality of Medical Care? A Laboratory Experiment."
  Diese Arbeit ist 2013 als *Ruhr Economic Paper* No. 414 erschienen.