Testing the effects of remuneration schemes and social benefits on doctors' effort: A real effort laboratory experiment

WORKING PAPER 5

September 2014

Mylène Lagarde*, Duane Blaauw**

* Department of Global Health and Development, London School of Hygiene & Tropical Medicine, UK

** Health Policy Unit, University of the Witwatersrand, South Africa

This paper has been funded by UK aid from the UK Government, however the views expressed do not necessarily reflect the UK Government's official policies.

About RESYST

RESYST is a 5 year international research consortium which aims to enhance the resilience and responsiveness of health systems to promote health and health equity and reduce poverty. We conduct our research in several low and middle-income countries in Africa and Asia, seeking to identify lessons that are transferable across contexts.

Research focuses on three critical health systems components:

- **Financing**: focusing on how best to finance universal health coverage in low and middle-income countries.
- **Health workforce**: identifying effective, practical interventions to address human resource constraints.
- **Governance**: studying the relationships among frontline actors and mid-level management, and leadership in health policy implementation processes.

More information: http://resyst.lshtm.ac.uk

Corresponding author

Mylène Lagarde:

Tel: +44 207 927 2090

Email: mylene.lagarde@lshtm.ac.uk

Contents

List of tables and figures

Executive summary

In this study, we investigate the effects of fee-for-service, capitation and salary, as well as the presence of benefits for patients, on the quantity and quality of effort provided by medical students in a laboratory setting. We designed a real-effort experiment, the first of its kind, to simulate the incentives and context surrounding the provision of health care services. We find that salary produces the lowest quantity of output, however our data do not show that FFS leads to a higher quantity of output than capitation. We find that the highest quality is achieved when participants are paid by salary, followed by capitation. We also find that there is less shirking under salary. Finally, the presence of patient benefits significantly increases the quality of performance, while it has no significant impact on quantity of output.

1. Introduction

The labour supply behaviour of health care providers is a significant determinant of the performance of health care systems. When health care is provided through the public sector, governments can design the compensation structure of health care providers to influence their behaviours and ensure that health care expenditure is efficiently spent to purchase an optimal quantity and quality of care.

In the health economics literature, the incentives created by the three traditional compensation schemes to pay doctors – fee-for-service (FFS), salary, and capitation (CAP) – have been well described in theory (McGuire, 2000). Under FFS payment where the doctor is reimbursed for each item of service provided (the fee usually depends on the type of service), if the FFS rate exceeds the marginal cost of delivering additional services doctors are encouraged to deliver more services and over-serve patients. Under salary systems, doctors are paid a regular salary to work a specified number of hours, so there is no incentive to exert effort. Finally, under CAP contracts, doctors receive a fixed payment per patient (registered with them) per period, independent of the services provided. Such contracts provide an incentive to increase the number of patients enrolled but conditional on that, they encourage providers to limit the quantity of care provided, for example by encouraging the provision of preventive care to limit future visits of sick patients. Furthermore, if they offer a uniform reimbursement rate for patients with different risk profiles, CAP contracts will encourage providers to avoid more risky patients ("cream-skimming").

However, two additional problems complicate matters in the health care setting. First, the output of doctors is not one-dimensional, the quality and quantity of care provided make competing demands on doctors' time. Not only do doctors decide how many patients they see and how much health care services they provide to each patient, but also how responsive they are to patient needs, how much time they dedicate to each patient to listen to their concerns, take the time to make a thorough examination, identifying a correct diagnosis, etc. Because quality of care is typically much more difficult to measure, one would expect doctors to neglect it at the expense of quantity under high-powered incentives schemes such as FFS which focus on quantity of care (Holmstrom and Milgrom, 1991). By contrast, low powered incentive schemes not linked to the quantity of services provided, such as salary and capitation, should reduce the detrimental impact on quality of care. Second, it has long been argued that social preferences play a role in the agency relationship between doctors and patients (Arrow, 1963). Because doctors care for their patients' welfare, the marginal cost of effort spent to provide good quality medical services is reduced because patients' benefits increased by higher quality of care.

While compensation policies have critical implications for the provision of health care, it has been challenging to study their effects empirically, and existing evidence is generally weak and inconclusive (Gosden et al., 1999; Scott et al., 2011). To study the effects of remuneration mechanisms on quality and quantity of care, the ideal data set would combine good measures of doctors' performance in a context where doctors are randomly allocated to different payment methods. Such data have proved virtually impossible to obtain in the field for several reasons. First, randomised controlled trials of remuneration schemes are difficult to implement in the health sector. Randomising health care providers to different remuneration schemes is politically unfeasible in most settings, and therefore experiments have relied on the voluntary participation of providers into a new ‘treatment’ group where a new financial scheme is introduced (to be compared to an existing one), introducing self-selection bias. Moreover, in the absence of experimental evidence, most quasi-experimental data are potentially fraught with the confounding effect of complex contextual factors and parallel interventions. Health care systems display complex interactions between payers, providers and patients that are likely to influence the behaviours of providers. For example, there is ample evidence suggesting that various patients characteristics (type of insurance coverage, socio-economic characteristics) influence clinical decision-making and quality of care (Bernheim et al., 2008; Fiscella et al., 2000). This issue makes it difficult to draw lessons from empirical studies carried out in different countries where the institutional settings are likely to drive differences as much as the remuneration schemes (Gosden et al., 1999). Furthermore, performance outcomes are either biased by measurement errors (e.g. relying on self-reported behaviour) or virtually impossible to observe. This is particularly true for quality of care, where definitional challenges have led to the utilisation of a variety of ways to measure quality from patients’ reported satisfaction (van Campen et al., 1995), average patient contact time (Gravelle et al., 2013), clinical observations (Leonard and Masatu, 2010) or proxy health outcomes (such as 30-day acute myocardial infarction mortality rate for hospital care) (McClellan and Staiger, 1999). Yet little consensus exists as to how quality can be accurately measured in health care, given its multi-dimensionality (Donabedian, 1966). Finally, even the best-designed field experiment would not be able to assess whether social preferences are likely to affect doctors’ behaviours and mitigate the effects of financial incentives. Indeed, it is not possible to create a counterfactual without patients benefit in the field, and it is also extremely challenging to measure patients’ benefits.

Because of these issues, laboratory experiments offer a valuable tool for studying remuneration mechanisms and social preferences in a controlled environment, complementing the evidence generated by observational field studies. In the laboratory, the experimenter defines precisely the material incentives upon which subjects base their effort decision. Since the experimenter also designs the effort task, it is possible to create one that allows the measurement of different dimensions of individual effort (e.g. quality and quantity of output). Following a call to apply experimental economics tools to the field of health (Fuchs, 2000), a few experimental studies have recently emerged that have looked at doctors’ behaviours in the lab, with a particular interest in the relative effects of different remuneration mechanisms (Brosig-Koch et al., 2013; Godager and Wiesen, 2013; Hennig-Schmidt et al., 2011; Keser et al., 2013).

This paper makes four major contributions to this emerging literature.

First, unlike all existing studies that have adopted a “chosen effort” approach, we use a real effort experiment to test the impact of the three main remuneration schemes (CAP, FFS and salary) on providers’ effort. In experimental economics, two approaches have been used to test the impact of different types of incentives and remuneration mechanisms on individual performance. In chosen effort experiments (e.g. (Bull et al., 1987; Fehr et al., 1993; Nalbantian and Schotter, 1997), participants are asked to choose a level of (hypothetical) effort, for which they will be remunerated according to a specified rate and method. In real effort experiments, participants are asked to perform actual simple tasks (for example counting letters (Rosaz and Villeval, 2012), making simple additions (Dohmen and Falk, 2011), or entering data (Greiner et al., 2011). Evidence suggests that people react similarly to incentives in the two approaches (Bruggen and Strobel, 2007), but it has been argued that real effort tasks are closer to real-life settings where work involves actual effort, boredom, and intrinsic motivation, which are not experienced when allocating budgets in chosen effort experiments (van Dijk et al., 2001). Since intrinsic motivation is potentially important in the context of health care, having a real effort task might be preferable.

Second, we are able to evaluate simultaneously the impact of remuneration schemes on two dimensions: quality and quantity of output. This is particularly relevant to the health care context, where a lot of policy and academic debates focus on the impact of payment mechanisms not only on the level of quantity of care provided but also on its quality. Unlike previous health economics experiments that have only focused on a unique measure of effort provided, in our experiment we include two dimensions and acknowledge that they are substitutes as far as the worker is concerned: the time spent to enter data more accurately is at the expense of entering a higher quantity of data.

Third, we explicitly evaluate the extent to which the presence of benefits to patients has an impact on the decisions made by doctors working under different remuneration schemes. Indeed, even before behavioural economics provided compelling evidence of the influence of other-regarding preferences, economists proposed that physicians could be altruistic (Arrow, 1963). However, the introduction of altruism has unclear consequences for the predictions of theoretical models[1]. Our experimental design allows us to test and quantify the behavioural economic assumption that providers take patients' benefits into account when they make decisions about the quality and quantity of care they provide.

Fourth, the real effort task we developed allows us to test for the existence of cream-skimming under the different remuneration mechanisms, and in particular whether a carefully designed risk-adjusted CAP schemes is able to prevent that problem. This is a novel and particularly important feature of the experiment that was not feasible with the chosen effort experiment.

In this experiment, medical students are asked to enter medical test results from laboratory test reports into a computer. During the task, where their effort is measured both in terms of quantity and quality, subjects are paid in three different ways: for each individual entry made (FFS), for each laboratory result completed (CAP), and for a specific amount of time (salary). In addition to this within-subject design, participants are randomly allocated to one of two treatments. In a PATIENT BENEFIT treatment, participants' effort results in monetary gains for patients if and only if they enter test results correctly. In the other treatment ('NO PATIENT BENEFIT') subjects' effort only generates benefits for themselves, independently of the quality of their work. The aim of the experiment is to compare the relative effects of the three traditional payment mechanisms for doctors (salary, CAP, FFS) on the quantity and quality of effort, in the presence of benefits to patients or not.

We find that salary significantly reduces the quantity of output produced, while we do not find evidence that FFS or CAP payments result in different quantities of services. On the other hand,

[1] The extent to which doctors will increase the quality and quantity of services they provide depends on the relative strength of their altruism compared to their own self-interest on the one hand, and on their interpretation of the potentially complex relationship between patients' benefits and the quantity and quality of care provided.

salary leads to the highest quality of output and less shirking, and CAP leads to higher quality than FFS. Furthermore, we find that over-servicing is reduced with salary and CAP compared with FFS, and gaming is reduced when individuals are paid by salary compared with FFS. We find no evidence that an effort-adjusted CAP scheme (or salary) leads to more cream-skimming than FFS. Finally, in the presence of patient benefits, individuals increase the quality of their output under all three payment mechanisms, and they shirk and game less under FFS and CAP. However, there is no evidence of an impact of patient benefits on the quantity of output produced, probably due to the existence of large quantity-quality trade-offs.

This article is structured as follows. Section 2 presents the economic game and the experimental procedures. Section 3 provides the results of the 'NO PATIENT BENEFIT' treatment arm and Section 4 presents the results of the patient benefit treatment. Section 5 concludes.

2. An experimental approach to studying remuneration in health care

2.1. The medical effort task

The experiment consisted of four periods of a mundane working task. In each period participants were given hardcopy laboratory test reports containing a series of laboratory test results. The task was to type those results into an input mask at the computer screen. As a task, data entry is easy to explain, it is unlikely to be much distorted by personal variables such as ability or experience, and it guarantees some heterogeneity in productivity and effort.

There were two types of laboratory reports: expanded and basic ones (see Figure A1 in Appendix for an example of an expanded report). Basic reports included 14 individual test results[2] while expanded reports included 22 different individual test results[3]. During each data entry period of eight minutes, subjects had to enter data from a pile of fifteen reports (five basic ones and ten expanded ones) that had been sorted in ascending order of the report reference number, so that each subject was presented the forms in the same way but participants were free to enter data from reports in the order they wanted.

We used the software z-Tree (Fischbacher, 2007) to create a data entry mask and record subject performance. When the task started, a first screen would appear where a subject would have to enter the reference number of a report whose data they wanted to enter. A second screen with a data entry mask would then appear, that matched the basic or expanded report format that the subject sought to enter (see screen captures in Figures A2 and A3 in Appendix). Having entered

[2] Results were divided between Full Blood Count tests which included 8 individual results (red blood cell count, haemoglobin, haematocrit, mean corpuscular volume, mean corpuscular haemoglobin, mean corpuscular haemoglobin concentration, white blood cells count, platelet count) and Urea & Electrolytes tests that included 6 individual results (sodium, potassium, chloride, bicarbonate, urea, creatinine).

[3] In addition to the results in the basic reports, they included the 8 individual results of a Liver Function Test (total bilirubin, conjugated bilirubin, alanine aminotransferase, aspartate aminotransferase, alkaline phosphatise, total protein, albumin and globulin).

some or all of the individual test results, subjects could record their data entries by pressing a "Record data" button. Then, regardless of the quality of the data entry, a new screen appeared asking for the reference number of the next report they wished to enter[4]. At the end of the period, if time ran out while a subject was still entering numbers from a given form, all of these entries were not recorded in the system. Finally, every other form, the data entry mask would show that four individual results were already entered in the computer system (see Figure A4 in Appendix) and subjects were told that in such cases, it was unnecessary to enter the figures a second time. However, nothing prevented them from entering the results again if they wished.

We intended this task to parallel the main characteristics of the medical decision-setting, with a report representing an interaction with a patient (i.e. a consultation), during which a doctor is expected to exert some effort and perform a number of tasks (ask questions, take the patient's blood pressure, examine the patient, etc.). Having two types of reports imitates the fact that some patients are healthier than others and require less medical attention (basic reports require less effort). Allowing the subjects to re-order the laboratory reports makes it possible for them to start with easy reports first, as some doctors might be inclined to select health patients (a behaviour termed "cream-skimming"). Finally, having data already entered on the computer system introduces the possibility for subjects to perform some unnecessary act that is not required by the patient's medical condition and brings no benefit (over-servicing).

2.2. Design of the experiment

There were two dimensions to the experiment: a within and between-subject design.

For the within-subject component of the design, subjects took part in four consecutive data entry periods of eight minutes each, after taking part in a three-minute training session[5]. During the first three periods, participants were paid differently for each period, according to the following remuneration schemes:

- A fixed payment of R125[6] for the 8-minute period, paid to subjects independently of their performance, both in terms of number of numbers entered or how many were correct. This payment is similar to paying doctors a salary (*SAL*).
- A payment of R1 for each individual test result entered, whether accurate or not. This remuneration scheme is mimicking the fee-for-service (*FFS*) payment system, as subjects also receive R1 if the subject enters an individual test result that was already recorded in the system, as a doctor would be paid to make tests even though they were not necessary.
- Finally in the third payment mechanism, subjects are paid for each report validated, at a rate of R20 for an expanded report and R15 for a basic report, whether reports were fully complete, and independently of the number of accurate entries. This payment system is similar to a risk-

[4] Subjects could only enter results from a laboratory report once: if they entered the reference number of a report previously entered, they would get an error message.

[5] Subjects were given two sample reports to enter into the computer. The training session were designed to provide an opportunity for subjects to get familiar with the different aspects of the data entry task (data entry mask, how to enter data efficiently, etc.).

[6] The exchange rate around the time of the experiment was about ZAR 10 for USD 0.965.

adjusted capitation (CAP) system, where providers are paid more for patients demanding more medical attention and more effort from their doctors[7].

The rates of the three payment mechanisms were carefully chosen after we piloted the task with a few subjects. We used the average performance during the pilot (between 6 and 7 reports) to set the three rates in a way that would make the three mechanisms income neutral, and only differing, on average, in the incentives they create.

The order in which subjects were paid in the first three periods was randomised to control for possible learning and order effects. We made sure to obtain a balanced cross-over design within each treatment arm, with an equal number of subjects allocated to the six different order sequences of the three payment mechanisms[8]. At the end of the first three periods, participants were asked to choose which mode of payment they wanted to be used for the fourth and final period. In this paper, we do not include data from the fourth period.

For the between-subject aspect of the design, subjects were randomly assigned to one of two treatments. In the first treatment ('NO PATIENT BENEFIT'), participants' effort did not yield any monetary benefit to anyone but themselves. In the second treatment (PATIENT BENEFIT), subjects effort translated into actual benefits for patients. Specifically, after subjects were initially asked to choose from a list of five local charities[9], they were informed that for each individual test results they would enter correctly, the charity of their choice would receive R0.50 to be used to provide medical care to patients. Subjects were also told that entering an "unnecessary" test result (one that was already loaded in the system) would not benefit patients.

2.3. Experimental protocol

The experiment was run in a computer laboratory with third and fourth year classes of medical students from the Faculty of Health Sciences at the University of Witwatersrand (Johannesburg, South Africa), recruited through adverts posted on their intranet and leaflets distributed in their classes. On the day of the experiment, participants were given a show-up fee of R50 (USD4.83) and were randomly allocated to a workstation. Most of the instructions during the experiment were presented on the computer screen. At the very beginning however, the same experimenter throughout all sessions provided a brief introduction to the session and some general instructions about the data entry activity. Participants were informed that they would be remunerated differently in different periods, and that information about payment would be given on the screen.

---

[7] In the present case, although the rates do not equalise perfectly the marginal cost of effort per entry, they get close to it: for an expanded form, individuals receive R0.91 per entry and for a basic form they receive R1.07 (a rate higher by 18%). This small differential, together with the fact that short reports will be validated more quickly (with more certainty within the allotted time), leaves an incentive to complete short forms first, although it is less obvious than with a unique rate.

[8] In each of the two treatments analysed here, exactly 11 individuals were allocated to each one of the six sequences.

[9] In a preliminary pilot, a unique charity was chosen which led to some negative comments from participants in the debriefing sessions, about the choice of that particular charity that was caring for patients suffering from a condition that was already attracting a lot of attention and support. To avoid similar negative reactions, we introduced the possibility for participants to choose the charity they wanted, out of the following ones: Witwatersrand Hospice, SOS Children's villages, South African National Tuberculosis Association, Cancer Association of South Africa, Thusanani Children's Foundation.

They were also told that only one of the four periods would be randomly selected for the final payments. Instructions can be found in Appendix.

At the end of the fourth period, one participant was invited to throw a die to decide which period would be chosen for payment. Payouts were then calculated and made to each participant anonymously (in a sealed envelope only identified with their study number) after they had completed a short questionnaire to collect socio-demographic information.

132 students participated in the experiment, 66 in each of the two treatments. A session lasted approximately 60 minutes and on average participants earned R167 per session (in addition to the show-up fee of R50), and the average transfer to charities in the patient treatments was R53.3. In total, R8,764 (about USD 846) were transferred to the five charities. Each charity received a bulk payment once all sessions had been run, with the money earmarked to pay for patients' medical care.

Table 1 shows the main characteristics of the 132 participants. Subjects were nearly 22 year old on average, there was a majority of female participants, 53% described themselves as Black-African and 22% as white. Finally, slightly less than half of the participants were in fourth year (the rest being in third year).

TABLE 1: DESCRIPTIVE STATISTICS

Variable	Obs.	Mean	Std. Dev.	Min	Max
Male	132	0.409	0.492	0	1
Age	132	21.705	2.144	19	40
Black	132	0.530	0.500	0	1
White	132	0.220	0.414	0	1
Fourth Year	132	0.439	0.497	0	1

2.4. Performance measures

Performance of subjects is evaluated on three types of outcomes: quantity of output, quality of output and undesirable behaviours.

Quantity of output is measured by the number of items entered in total during a given period (i.e. number of acts performed to follow the medical analogy) and the number of forms completed (total number of patients seen in a given period). Furthermore, to capture over-servicing we create a variable capturing the total number of items unnecessarily entered (results that were already entered on the system but were then re-entered by subjects).

Quality of output is measured by four different variables. First, simply by the number of entries correctly made that were both correct and necessary[10] (the number of services correctly performed), second by a quality index calculated as the proportion of entries correctly made. This third indicator is created to detect. Finally, we construct a dummy variable to detect the propensity of subjects to process reports in a different order than the one in which they were presented, more specifically to prioritise basic reports first ("cream-skimming"). Lastly, we look at shirking behaviour, the number of forms for which less than 90% of entries are correctly made.

3. Testable hypotheses

Based on the theoretical insights about the three payment mechanisms, we can formulate a number of hypotheses that can be tested in the laboratory game.

Concerning the quantity of effort made, we formulate four hypotheses:

(1) FFS will yield the highest number of entries, as FFS provide high-powered incentive to increase that output.

(2) We also expect that over-servicing will be most prominent under FFS compared to the other two schemes for that reason.

(3) Under the CAP scheme, participants will seek to maximise the number of reports validated (for which they are paid), but not necessarily the number of individual entries made. This is similar to maximising the size of one's patient list.

(4) Because we hypothesise that subjects are altruistic, under the PATIENT BENEFIT treatment, we expect no impact on the quantity of output provided, due to the existence of two potential strategies. Some subjects might want to focus on the accuracy of their data entries, at the expense of the quantity of tests entered. Others might seek to increase benefits for patients by simultaneously reaching high levels of productivity and high level of accuracy.

Concerning the quality of output, we formulate the following three hypotheses:

(1) Quality of care will be lower with high-powered incentive schemes tying remuneration to a quantitative output (number of tests entered for FFS ; number of reports validated for CAP), compared to the low-powered incentive provided by salary.

(2) Because the experiment uses an "effort-adjusted" CAP remuneration, we do not anticipate that under the CAP system participants will be more likely to enter short reports first (cream-skimming).

(3) Finally, when patients benefit from a higher quality of services produced, we expect that our subjects, prospective doctors who self-selected into the medical profession, will confirm the altruistic physician hypothesis and make more effort to increase the quality of their work as they care for patients.

[10] We excluded unnecessary entries, i.e. those made by subjects even though the numbers had already been entered on the computer.

4. Results

Table 2 provides a summary of the summary statistics of all the performance measures used in the analysis.

Looking at the impact of incentives on quantity of effort (panel A), on average subjects achieved the highest quantitative performance under the FFS payment. This is particularly obvious if we consider the number of items entered. Looking across treatments, differences point to a reduction in effort under the PATIENT BENEFIT treatment, although these differences are not always marked.

With regards to the quality of performance (panel B), the highest performance seems to be achieved under the salary payment, where in absolute terms subjects enter a higher number of items correctly, and in relative terms, the quality of their work (controlling for the quantity of their output) is also better. It follows that shirking is also more common when subjects are paid by CAP or FFS. Unlike results measured by the quantity of effort in Panel A, comparisons of the quality of output exerted between the two treatments point to a clear improvement under the PATIENT BENEFIT treatment, where all indicators improved: there was a higher number of items entered correctly, a higher quality index and a reduction in shirking. Finally, with regards to cream-skimming, there is no clear pattern emerging across remuneration schemes and treatments. By contrast, it seems that of all remuneration mechanisms, only CAP led to a clear preference for more work-intensive patients, even in the PATIENT BENEFIT treatment, even though differences were slightly less marked.

TABLE 2: DESCRIPTIVE STATISTICS OF PERFORMANCE UNDER THE TWO TREATMENTS

	NO PATIENT BENEFIT treatment		PATIENT BENEFIT treatment	
	mean	SD	mean	SD
Panel A: quantity of performance				
Number of items entered				
SAL	147.73	57.00	144.67	46.55
FFS	199.36	79.65	179.15	72.30
CAP	188.09	77.20	171.12	69.56
Number of forms processed				
SAL	8.64	3.08	8.03	2.37
FFS	10.68	3.80	9.64	3.58
CAP	10.36	3.78	9.56	3.57
Over-servicing [a]				
SAL	7.58	7.41	7.94	7.45
FFS	15.14	10.17	12.21	9.32
CAP	11.74	10.61	9.62	8.77
Panel B: quality of performance				
Number of items correctly entered				
SAL	101.11	49.54	118.36	39.42
FFS	64.95	55.17	93.65	52.94
CAP	77.95	58.54	104.71	45.10
Accuracy rate [b]				
SAL	0.75	0.34	0.86	0.21
FFS	0.47	0.42	0.63	0.35
CAP	0.56	0.41	0.71	0.31
Shirking behaviour [c]				
SAL	3.70	4.77	2.27	3.00
FFS	7.83	5.97	5.41	5.19
CAP	6.83	5.83	4.67	4.64
Proportion of cream-skimmers [e]				
SAL	0.15	0.36	0.14	0.35
FFS	0.17	0.38	0.11	0.31
CAP	0.17	0.38	0.14	0.35

[a] Total number of items entered unnecessarily by the subject (the correct numbers were already entered).

[b] Proportion of items entered correctly.

[c] Number of poorly processed report, defined as a report where less than 90% of items were entered correctly.

[d] Proportion of subjects who purposefully re-sorted the reports to start with the basic ones first.

In the analysis below, we consider the two types of outcomes (quantity and quality) separately. First, we look at the causal impact of physicians' remuneration schemes on performance in the absence of benefits to patients. This is done by analysing the data from the 'NO PATIENT BENEFIT' treatment only. Then, we pool the data from the two treatments together to evaluate the causal impact of the presence of patient benefits on performance, and investigate whether that effect varies across remuneration schemes.

4.1. Effects of remuneration schemes on physicians' performance

4.1.1. Impact on quantity of effort exerted

We use two indicators for the quantity of output produced: the number of individual entries and the number of reports completed.

First, we carry out a within-subject analysis of the BASE treatment data, taking advantage of the fact that each subject was paid differently in each of the three periods of the game[11]. To identify the effect of salary and CAP payment schemes compared to FFS, we regress the outcomes of interest on two dummies for salary and CAP. To capture within-individual effects, we use fixed-effects linear regressions.

Table 3 shows that salary induces less effort than FFS, but there is no evidence of a difference between CAP and FFS: subjects who were paid by salary entered fewer numbers (columns 1-2, p=0.000 with controls for period), and completed more reports (columns 3-4, p= 0.007 with period controls). On average, when they were paid by salary individuals made nearly 52 fewer entries than when paid by FFS, a reduction in productivity by slightly more than 25%. Looking at the number of reports completed during a period, we find that when subjects received a salary, they saw about two patients less than what they did when paid by FFS, corresponding to a reduction in output by slightly less than 20%[12].

Finally the results show consistent evidence that salary and CAP, which do not incentivise individual items, lead to less over-servicing compared to FFS. On average, being paid by salary and CAP both led to a reduction by 15% in the number of unnecessary entries made (2.2 entries) compared to being paid by FFS. This result is robust to the inclusion of period effect controls (column 6)[13].

[11] This analysis is possible because order of the three remuneration schemes was randomised across subjects.

[12] The small difference between the two outcomes is likely to be driven by an end-of-game effect: as they run out of time at the end of the period, subjects would have validated their entries in an incomplete form (therefore not fully completing and validating the final form).

[13] Interestingly, we do not find evidence that the reduction in over-servicing seems is different for both salary and capitation (Wald test, p<0.986), while descriptive evidence in Table 1 suggested that salary led to a lower reduction.

TABLE 3: FIXED-EFFECTS LINEAR REGRESSIONS OF THE IMPACT OF REMUNERATION SCHEMES ON QUANTITY OF EFFORTS IN THE 'NO PATIENT' TREATMENT

	Number of items entered		Number of reports completed		Over-servicing	
	(1)	(2)	(3)	(4)	(5)	(6)
SAL	-51.636***	-51.636***	-2.045***	-2.045***	-2.163**	-2.213**
	(10.235)	(9.591)	(0.619)	(0.590)	(0.763)	(0.846)
CAP	-11.273	-11.273	-0.318	-0.318	-2.216***	-2.226***
	(10.679)	(9.291)	(0.468)	(0.399)	(0.551)	(0.537)
Period 2		22.455**		1.061**		0.159
		(7.007)		(0.332)		(0.856)
Period 3		22.182***		1.212**		0.400
		(5.696)		(0.376)		(1.105)
					0.105***	0.104***
					(0.006)	(0.008)
Constant	199.364***	184.485***	10.682***	9.924***	-5.702***	-5.698***
	(6.536)	(5.011)	(0.336)	(0.360)	(1.313)	(1.293)
Observations	198	198	198	198	198	198
R^2	0.255	0.312	0.209	0.284	0.698	0.698

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

4.1.2. Impact on quality of output

We now turn to the impact of remuneration mechanisms on the quality of output in a given period. Table 4 presents the results of fixed-effects linear regressions assessing the effect of payment schemes on three indicators of the quality of output in the 'NO PATIENT BENEFIT' treatment of the medical game. Considering the potential trade-off between quality and quantity of output, we include controls for the quantity of output produced by each subject during the period of interest.

Table 4 shows that salary and sometimes CAP induces better quality of work than FFS: subjects who were paid by salary and CAP made more correct entries (columns 1-2, p=0.086 for CAP and p=0.010 for salary with controls for period) than those paid by FFS; salary had also a positive impact on the quality index of production undertaken (columns 3-4, p=0.011 with period controls) and yielded less shirking than FFS (columns 5-6, p=0.014 with period controls). But when looking at the quality index (columns 3-4, p=0.128 with period controls) or the incidence of shirking (columns 5-6, p=0.148 with period controls), there was no evidence of a difference in quality of output under CAP and FFS.

On average, subjects paid by salary made about 15 more correct entries compared to when they were paid FFS, an increase by slightly more than 23%. When paid by CAP, subjects performed better than under FFS by about 12.3% on average (about 8 more correct entries). Despite an apparent difference between salary and CAP, we do not find evidence of a differential impact of salary and

CAP (Wald test, p<0.265). Turning to the impact on an index of quality of output produced, we find that when subjects received a salary, their accuracy was better than under FFS by 10 percentage points, corresponding to an improved performance of 20%. In other words, they were about 10% more accurate when entering numbers. For the final indicator measuring the number of forms where shirking occurred, we find that when paid by salary, subjects shirked their effort less often, more specifically this behaviour occurred 15% less frequently (in slightly less than 1.2 reports).

TABLE 4: FIXED-EFFECTS LINEAR REGRESSIONS OF THE IMPACT OF REMUNERATION SCHEMES ON QUALITY OF OUTPUT IN THE 'NO PATIENT BENEFIT' TREATMENT

	Number of correct entries		Accuracy rate		Shirking behaviour	
	(1)	(2)	(3)	(4)	(5)	(6)
SAL	15.868***	14.613***	0.095***	0.096**	-1.225***	-1.144**
	(4.276)	(4.490)	(0.027)	(0.030)	(0.358)	(0.376)
CAP	8.572*	8.298*	0.045	0.046	-0.547	-0.534
	(4.536)	(4.304)	(0.026)	(0.027)	(0.351)	(0.338)
Period 2		3.700		-0.007		-0.203
		(5.316)		(0.031)		(0.238)
Period 3		10.525		0.001		-0.667
		(7.363)		(0.044)		(0.396)
Entries made	-0.393**	-0.417**	-0.004***	-0.004***		
	(0.128)	(0.135)	(0.001)	(0.001)		
Completed reports					1.423***	1.463***
					(0.139)	(0.122)
Constant	143.267***	143.372***	1.186***	1.185***	-7.370***	-7.505***
	(24.047)	(24.585)	(0.161)	(0.162)	(1.468)	(1.351)
Observations	198	198	198	198	198	198
R^2	0.403	0.419	0.617	0.618	0.786	0.793

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

On the other hand, we find mixed evidence on the effect of CAP on quality of output. The small positive effect found on the number of entries made suggests that the incentives created by CAP were qualitatively different from those created by FFS. In other words, output-based payments bundling services together are likely to incentivise individuals to focus on quality more than itemised payments. Yet the absence of impact on the quality index and shirking behaviour suggests that this positive effect might be limited.

Finally, there is no evidence supporting the idea that subjects displayed more cream-skimming behaviour under CAP, compared to FFS or salary (Table 5). This result confirms the hypothesis that well-adjusted CAP payment does not necessarily lead to prioritising easy cases before more difficult ones.

TABLE 5: FIXED-EFFECTS REGRESSIONS OF THE IMPACT OF REMUNERATION SCHEMES ON UNDESIRABLE BEHAVIOURS IN THE 'NO PATIENT BENEFIT' TREATMENT

	1= put shorter forms first	
	(1)	(2)
SAL	-0.186	-0.243
	(0.969)	(0.989)
CAP	0.828	0.782
	(0.892)	(0.908)
Period 2		-0.465
		(0.856)
Period 3		-0.754
		(0.916)
Entries made	-0.013*	-0.012
	(0.008)	(0.008)
Completed reports		
Constant	-2.402	-2.221
	(1.665)	(1.697)
Observations	198	198
R^2	-41.59	-41.21

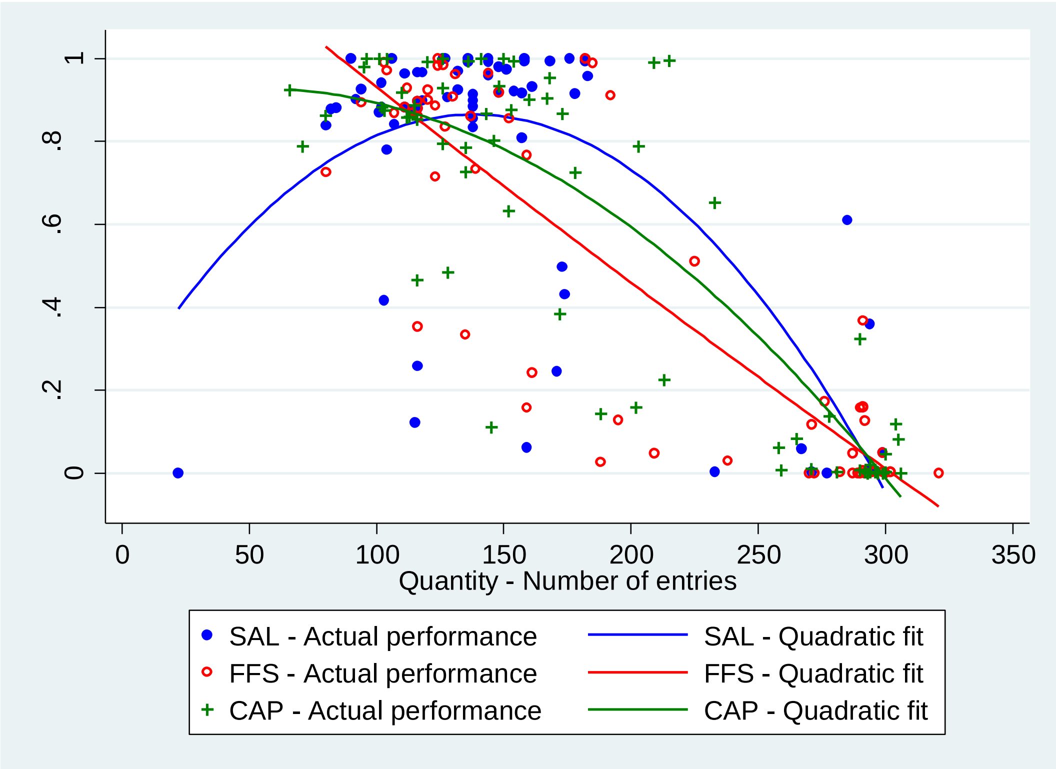
Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

4.1.3. Quantity and quality trade-offs

In all the specifications presented in Table 4, the coefficient associated with the variable controlling for quantity of effort is negative and significant. This clearly demonstrates the existence of a quantity-quality trade-off in the effort exerted in the game, as there are in the production of medical services.

To further investigate the nature of trade-offs under each payment mechanism, we plot the performance of respondents along the two main dimensions (quality index vs. different levels of quantity of output), and show the results in Figure 2.

Figure 1: Quantity-quality trade-offs in the 'NO PATIENT' treatment



Several results emerge. First, unsurprisingly, for all payment mechanisms the marginal rate of substitution between quality and quantity is the highest at high levels of output. In other words, in order to reach high levels of output in a limited amount of time, quality has to be sacrificed. Second, there is a clear difference between the three payment mechanisms regarding the level at which that sacrifice starts. Under salary, quality starts being sacrificed for higher outputs[14]. In fact, the absence of focus on output levels under the salary payment guarantees a higher quality at higher level of outputs. Third, the bell shape of the salary curve in the salary treatment suggests that some subjects' behaviour follows the theoretical predictions and make very little effort both in terms of quantity and quality.

[14] If we regress the accuracy rate on the number of entries and number of entries squared, we find that negative returns on quality are reached at an output level of about 136 entries under salary, for CAP it is around 53 entries and for FFS the negative relationship between quality and quantity exists right from the beginning.

4.2. Effects of the presence of patients' benefits on physicians' performance

Here we use both the within-subject and between-subject nature of the data, to formally compare subjects' performance in the two treatments, and investigate the heterogeneity of that effect across remuneration schemes.

4.2.1. Impact on quantity of effort exerted

The results on the impact of the presence of benefits to patients on productivity, presented in Table 6, show that subjects did not increase the quantity of output they produced. Indeed the coefficient associated with the PATIENT BENEFIT treatment is not significant, either when we look at the number of entries made (column 1, p=0.276), or the number of reports completed (column 4, p=0.170). Breaking down the effect of the treatment across the three payment mechanisms to test whether the treatment effect is different for different remuneration schemes, with or without period controls (columns 2-3 and 5-6), we also find no evidence that patient benefits have an impact on productivity.

Furthermore, we find no evidence supporting the idea that, even if patients do not benefit from the provision of unnecessary services, there is no reduction of over-servicing (column 7-9).

TABLE 6: RANDOM-EFFECTS LINEAR REGRESSIONS OF THE IMPACT OF REMUNERATION SCHEMES ON QUANTITY OF EFFORT

	Number of items entered			Number of reports completed			Over-servicing		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SAL	-43.061***	-51.636***	-51.636***	-1.826***	-2.045***	-2.045***	-1.967***	-2.860***	-2.832***
	(8.426)	(10.248)	(9.520)	(0.488)	(0.620)	(0.586)	(0.375)	(0.825)	(0.834)
CAP	-9.652	-11.273	-11.273	-0.197	-0.318	-0.318	-2.107***	-2.368***	-2.362***
	(6.958)	(10.692)	(9.140)	(0.292)	(0.469)	(0.392)	(0.412)	(0.581)	(0.558)
PATIENT	-13.414			-0.818			-0.330		
	(12.317)			(0.596)			(0.820)		
PATIENT*FFS		-20.212	-20.212		-1.045	-1.045		-1.084	-1.073
		(15.742)	(15.687)		(0.714)	(0.709)		(1.041)	(1.061)
PATIENT *SAL		-3.061	-3.061		-0.606	-0.606		0.642	0.644
		(11.320)	(11.869)		(0.585)	(0.615)		(0.849)	(0.848)
PATIENT *CAP		-16.970	-16.970		-0.803	-0.803		-0.576	-0.567
		(13.342)	(12.572)		(0.684)	(0.639)		(1.194)	(1.199)
Period 2			25.902***			1.295***			-0.508
			(6.700)			(0.318)			(0.405)
Period 3			24.773***			1.386***			0.186
			(3.945)			(0.180)			(0.750)
							0.092***	0.091***	0.092***
							(0.005)	(0.005)	(0.005)
Constant	195.965***	199.364***	182.472***	10.568***	10.682***	9.788***	-3.520***	-3.011***	-3.012***
	(18.320)	(19.910)	(19.178)	(0.883)	(0.964)	(0.956)	(0.736)	(0.968)	(0.880)
Observations	396	396	396	396	396	396	396	396	396
R^2	0.211	0.220	0.308	0.189	0.191	0.304	0.574	0.578	0.581

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

4.2.2. Impact on quality of effort

Pooling data from the two treatments, we now turn to the impact of patient benefit on the quality of output. Table 7 presents the results of random-effects linear regressions assessing the impact of the presence of patient benefits on the three quality indicators. As before, given the potential quality-quantity trade-off, we include controls for the quantity of output produced.

Overall, the results show that when the quality of their work has a direct benefit for patients, subjects work better: they increase the absolute number of correct entries they make (column 1, p=0.002), they improve the accuracy of their entries (column 4, p=0.001) and they shirk less (column 7, p=0.012). The results also show consistent evidence that the effect of patient benefit is significant and of the same magnitude for all types of payment schemes, except in one case: there is no evidence that patient benefit decreases shirking behaviour when subjects receive a salary (column 9, p=0.283).

On average, the existence of benefits to patients led to an increase of nearly 20 correct entries. Breaking down that effect for each of the different incentive schemes (and controlling for period effect, column 3), we find an increase of slightly more than 21 entries under FFS (an improvement by 33%), 16 entries under salary (an improvement by 16%) and about 21 entries under CAP (an improvement by 26%). We do not find evidence that these effects are significantly different in magnitude (Wald test, p<0.862). Turning to the overall quality of output produced, we find that when subjects knew their actions had a benefit for patients, their accuracy was better by about 9 percentage points (column 4). Looking at the impact for each payment mechanism (columns 5 and 6), we see that this effect amounts to about 8 percentage points under FFS (an increase by about 16%), to 10 percentage points under salary (an increase by about 13%) and to 9 percentage points under CAP (an increase by about 16%). There is no evidence that these effects are significantly different from one another (Wald test, p<0.867).

Finally, we find that on average the existence of patient benefits leads to a reduction in shirking: subjects shirked their effort less often, on average in 1 less report (column 7). More specifically, there was a reduction in shirking by 16% under CAP (1.11 reports) and 13.5% under FFS (1.05 reports). However, there is no evidence that patient benefit decreases shirking behaviour when subjects receive a salary. This is probably explained by the fact that shirking under salary is a rare phenomenon in the first place: it occurs more than half as much when subjects are paid by salary compared to the two other compensation schemes.

Finally, as in the NO PATIENT BENEFIT treatment, there is no evidence of an increase in cream-skimming under an effort-adjusted CAP scheme under the PATIENT BENEFIT treatment (See Table A1 in appendix).

TABLE 7: RANDOM-EFFECTS LINEAR REGRESSIONS OF THE IMPACT OF REMUNERATION SCHEMES ON QUALITY OF OUTPUT, POOLED DATA

	Number of correct entries			Accuracy rate			Shirking behaviour		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SAL	16.089***	19.100***	17.482***	0.092***	0.082***	0.081**	-1.298***	-1.520***	-1.456***
	(1.991)	(5.463)	(5.731)	(0.016)	(0.032)	(0.032)	(0.160)	(0.345)	(0.364)
CAP	8.816***	9.277**	8.924**	0.048***	0.043*	0.042*	-0.619***	-0.593*	-0.583*
	(3.137)	(4.019)	(3.748)	(0.016)	(0.025)	(0.024)	(0.177)	(0.322)	(0.316)
PATIENT	19.769***			0.088***			-0.957**		
	(6.240)			(0.026)			(0.379)		
PATIENT*FFS		22.022***	21.389***		0.078**	0.077**		-1.087***	-1.054***
		(7.769)	(7.869)		(0.034)	(0.033)		(0.333)	(0.336)
PATIENT *SAL		16.247*	16.151*		0.097**	0.097**		-0.649	-0.630
		(8.777)	(9.064)		(0.041)	(0.041)		(0.571)	(0.587)
PATIENT *CAP		21.154***	20.622***		0.088***	0.088***		-1.139**	-1.114**
		(6.632)	(6.475)		(0.028)	(0.027)		(0.487)	(0.486)
Period 2			5.183*			-0.004			-0.159
			(2.974)			(0.016)			(0.145)
Period 3			13.282***			0.018			-0.635**
			(3.912)			(0.025)			(0.264)
Entries made	-0.333***	-0.330***	-0.362***	-0.004***	-0.004***	-0.004***			
	(0.070)	(0.067)	(0.070)	(0.000)	(0.000)	(0.000)			
Completed reports							1.281***	1.279***	1.310***
							(0.066)	(0.064)	(0.062)
Constant	132.455***	130.790***	130.880***	1.231***	1.237***	1.236***	-5.914***	-5.830***	-5.898***
	(10.982)	(9.406)	(10.011)	(0.055)	(0.050)	(0.050)	(0.652)	(0.564)	(0.559)
Observations	396	396	396	396	396	396	396	396	396
R^2	0.407	0.408	0.448	0.684	0.685	0.687	0.787	0.788	0.797

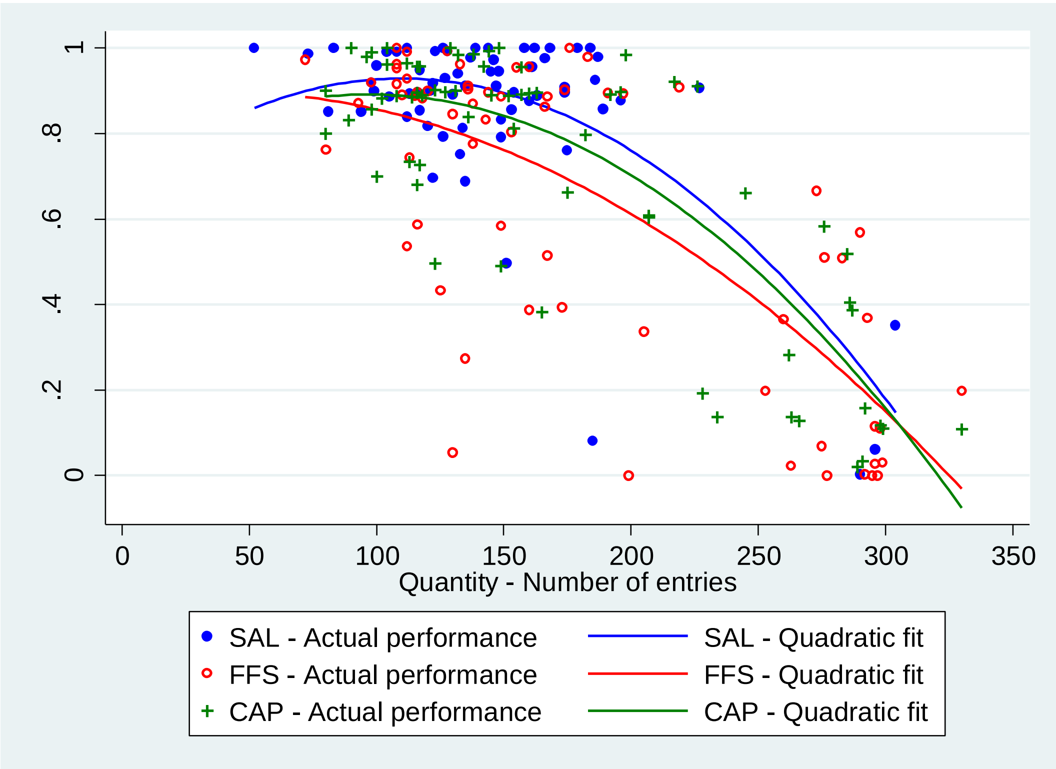
Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

4.2.1. Quantity and quality trade-offs

As in the previous treatment, we further look at the quantity-quality trade-offs in the effort exerted in the 'PATIENT BENEFIT' treatment. As before, we plot the performance of respondents under each remuneration mechanism, along the two main dimensions (quality index vs. different levels of quantity of output), and show the results in Figure 3.

Unlike the stark differences observed in the other treatment, under the PATIENT BENEFIT treatment, behaviours converge, with altruistic behaviour towards patients driving doctors to focus more on quality, especially at higher levels of production. It should be noticed that under salary, the presence of patient's benefits is particularly effective at increasing the level of effort at lower levels of productivity. The comparison between results under the two treatments for each remuneration scheme separately is more obvious in the graphs reproduced in Figure A5 in Appendix.

Figure 2: Quantity-quality trade-offs in the 'PATIENT' treatment



4.3. Cost-benefit analysis

Based on the total cost and performance outcomes under each of the three remuneration schemes, it is easy to compute the benefit-to-cost ratios for each of the remuneration schemes.

TABLE 8: BENEFIT-COST RATIOS UNDER EACH REMUNERATION SCHEME

	Salary	CAP	FFS
PANEL A: Productivity (number of test results entered)			
'NO PATIENT BENEFIT' treatment			
Average payment to subject (in ZAR)	125.00	190.75	199.36
Average number of items entered	147.73	188.09	199.36
Benefit cost ratio	**1.18**	**0.99**	**1.00**
'PATIENT BENEFIT' treatment			
Average payment to subject (in ZAR)	125.00	176.52	179.15
Average number of items entered	144.67	171.12	179.15
Benefit cost ratio	**1.16**	**0.97**	**1.00**
PANEL B: Quality (number of correct test results)			
'NO PATIENT BENEFIT' treatment			
Average number of correct test results	118.36	104.71	93.65
Benefit cost ratio	**0.95**	**0.59**	**0.52**
'PATIENT BENEFIT' treatment			
Average number of correct test results	101.11	77.95	64.95
Benefit cost ratio	**0.81**	**0.41**	**0.33**

The results presented in Table 8 show that for all types of outcomes, and for both treatments, the option with the highest benefit-cost ratio is always the salary scheme. In the 'NO PATIENT BENEFIT' treatment, we find that salary yields 1.18 entries per monetary unit, as opposed to one entry under the other two payment mechanisms. Salary is also the better option in the 'PATIENT BENEFIT' treatment, where the ratios are very similar to the ones observed in the other treatment.

Looking at quality of the output, we find that benefit-cost ratios are almost twice as big with salary compared to the other two remuneration schemes, in both treatments. Interestingly, the benefit-cost ratios are higher for all three payment mechanisms in the presence of benefits for patients, due higher levels of performance.

5. Discussion and conclusions

In this paper, we set out to evaluate the effects of the three main remuneration schemes used to pay individual doctors (salary, CAP and FFS) on the level of productivity and quality of effort provided. We also sought to test the impact on these outcomes of the altruistic doctor hypothesis. To isolate the effects of the three incentive schemes and the presence of benefits to patients, we developed a real effort experiment that intended to reproduce the main aspects of clinical decision making, while allowing us to exogenously change the different incentives influencing subjects' decisions and measure precisely the different outcomes of interest.

We found that salary led to the lowest level of productivity, compared to FFS and CAP. Specifically, under a salary payment, individuals focused more on the quality of their output than under an output-based payment mechanism. This result matches previous findings from the experimental and labour economics literature (Greiner et al., 2011; Lazear, 2000), but to some extent it has been less discussed in the health policy arena. In fact, in primary care, doctors are rarely remunerated through a salary, except in low- and middle-income countries, where the levels of their remuneration is typically very low and seen as responsible for low level of performance (McCoy et al., 2008). To some extent our results, obtained in conditions designed specifically to make the three mechanisms income neutral and leave only a role to the incentive structure created by the remuneration mechanism, reminds that, everything else being equal (and in particular for an equivalent average remuneration), salaries are potentially a better solution to preserve quality of care. Interestingly, we showed that with a CAP scheme well adjusted for the quantity of effort produced, cream-skimming does not occur.

Finally, we found compelling evidence that (prospective) doctors care about patients' benefits to the extent that this altruism contributes to mitigating the incentives created by payment mechanisms. Specifically, in our experiment prospective doctors made more effort irrespective of the payment scheme used, and shirking decreased under FFS and CAP where they were the most prevalent. In other words, we found that when the link between doctors' effort and patients' benefit is clear and that doctors are aware of it, doctors' altruism contributes to mitigating the incentives created by remuneration mechanisms in a direction that results in greater benefit-cost ratios. Of course, in the laboratory experiment the existence of the patient's benefit is not only obvious, it is clearly quantified and therefore evident to doctors. In real life, this link is not necessarily obvious and/or present on providers' mind for every single act or interaction they have with patients. It is also hardly ever quantified (except if doctors have in mind the QALY benefits of a particular intervention based on clinical trial results). Our results suggest that, in settings where doctors are not exerting enough effort in their job, policy-makers could remind them of the benefits of their actions (or lack thereof) for patients, in an attempt to make the benefits of their effort more salient.

Finally, thanks to precise measures of quantity and quality of outcomes produced, we were able to provide evidence of the existence of large quantity-quality trade-offs, highlighting the fundamental conundrum in the delivery of health care services. These trade-offs, clearly identified in the experiment here, are not often discussed or studied in the real life, probably due to challenges involved in collecting adequate data. This experimental setting is particularly helpful in reminding researchers and policy-makers alike of the importance to remember their existence and refrain from focusing too much on one dimension of performance while ignoring the other one. To some extent, whilst certain recent policy initiatives such as pay-for-performance seek to focus providers' effort on the quality of care they provide, very little research or attention is given to the impact of such initiatives on the quantity of services provided.

More than a decade after Fuchs' call (Fuchs, 2000), this work contributes to the early development of what can be considered as a promising area of work. Surprisingly, very little work has been done using the tools developed in experimental economics to look at the critical questions raised in the health care sector. While economists working in other fields (labour economics, industrial economics, etc.), have already embraced this approach, health economists remain more reluctant to the abstraction of laboratory experiments. While the external validity of economic experiments outside the laboratory is obviously questionable, the results obtained in a controlled environment can be useful to think through the complex challenges posed by the complex interaction of actors (providers, patients, payers) and the incentives (various remuneration schemes, contracts and social preferences) emerging from their relationships. It is hoped that this paper will add to this nascent literature by providing a new approach to studying remuneration schemes and modelling the medical decision making environment. Extensions of this work include the revision of the CAP remuneration scheme to make the incentives of that scheme (i.e. initial choice of patients) slightly more salient. Testing the impact of interventions designed to improve the quality of care provided (such as Pay-for-Performance and public reporting) has also been a natural extension of this work, which will be reported in another paper.

References

1. Arrow K. Uncertainty and the Welfare Economics of Medical Care. *American Economic Review* 1963;53; 941-943.
2. Bernheim SM, Ross JS, Krumholz HM, Bradley EH. Influence of patients' socioeconomic status on clinical management decisions: a qualitative study. *Ann Fam Med* 2008;6; 53-59.
3. Brosig-Koch J, Hennig-Schmidt H, Kairies N, Wiesen D, How to Improve Patient Care? – An Analysis of Capitation Fee-for-Service, and Mixed Payment Schemes for Physicians. 2013. ((Eds), Ruhr Economic Papers #412 Ruhr-Universität Bochum (RUB), Department of Economics: Essen, Germany; 2013.
4. Bruggen A, Strobel M. Real effort versus chosen effort in experiments. *Economics Letters* 2007;96; 232-236.
5. Bull C, Schotter A, Weigelt K. Tournaments and Piece Rates: An Experimental Study. *Journal of Political Economy* 1987;95; 1-33.
6. Dohmen T, Falk A. Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. *American Economic Review* 2011;101; 556-590.
7. Donabedian A. Evaluating the Quality of Medical Care. The Milbank Memorial Fund Quarterly 1966;44; 166-206.
8. Fehr E, Kirchsteiger G, Riedl A. Does Fairness Prevent Market Clearing? An Experimental Investigation. *Quarterly Journal of Economics* 1993;108; 437-459.
9. Fiscella K, Franks P, Gold MR, Clancy CM. Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care. *Jama* 2000;283; 2579-2584.
10. Fischbacher U. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 2007;10; 171-178.
11. Fuchs VR. The future of health economics. Journal of Health Economics 2000;19; 141-157.
12. Godager G, Wiesen D. Profit or patients' health benefit? Exploring the heterogeneity in physician altruism. *J Health Econ* 2013;32; 1105-1116.
13. Gosden T, Pedersen L, Torgerso D. How should we pay doctors? A systematic review of salary payments and their effect on doctor behaviour. *Q J Med* 1999; 47-55.
14. Gravelle H, Scott A, Sivey P, Yong J. 2013. Competition, Prices, and Quality in the Market for Physician Consultations. In: York Uo (Ed)^(Eds). 2013.
15. Greiner B, Ockenfels A, Werner P. Wage transparency and performance: A real-effort experiment. *Economics Letters* 2011;111; 236-238.
16. Hennig-Schmidt H, Selten R, Wiesen D. How payment systems affect physicians' provision behaviour—An experimental investigation. Journal of Health Economics 2011;30; 637-646.
17. Holmstrom B, Milgrom P. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. J*ournal of Law, Economics, & Organization* 1991;7; 24-52.
18. Keser C, Montmarquette C, Schmidt M, Schnitzler C. 2013. Custom-Made Healthcare – An Experimental Investigation. (Ed)^(Eds), *CIRANO - Scientific Publications* 2013s-15 Gottingen, Germany; 2013.
19. Lazear EP. Performance Pay and Productivity. *American Economic Review* 2000;90; 1346-1361.

20. Leonard KL, Masatu MC. Using the Hawthorne effect to examine the gap between a doctor's best possible practice and actual performance. *Journal of Development Economics* 2010;93; 226-234.
21. McClellan M, Staiger DO. 1999. The Quality of Health Care Providers. (Ed)^(Eds), *NBER Working Paper* #7327. 1999.
22. McCoy D, Bennett S, Witter S, Pond B, Baker B, Gow J, Chand S, Ensor T, McPake B. Salaries and incomes of health workers in sub-Saharan Africa. *Lancet* 2008; 675-681.
23. McGuire TG. 2000. Physician agency. In: Culyer AJ, Newhouse JP (Ed)^(Eds), *Handbook of Health Economics*. North-Holland Elsevier: Amsterdam; 2000.
24. Nalbantian HR, Schotter A. Productivity Under Group Incentives: An Experimental Study. *American Economic Review* 1997;87; 314-341.
25. Rosaz J, Villeval MC. Les and Biased Evaluation: a Real-Effort Experiment. *Journal of Economic Behavior and Organization* 2012;84; 537-549.
26. Scott A, Sivey P, Ait Ouakrim D, Willenberg L, Naccarella L, Furler J, Young D. The effect of financial incentives on the quality of health care provided by primary care physicians. *Cochrane Database Syst Rev*. 2011; CD008451. doi: 008410.001002/14651858.CD14008451.pub14651852.
27. van Campen C, Sixma H, Friele RD, Kerssens JJ, Peters L. Quality of care and patient satisfaction: a review of measuring instruments. *Med Care Res Rev* 1995;52; 109-133.
28. van Dijk F, Sonnemans J, van Winden F. Incentive systems in a real effort experiment. *European Economic Review* 2001;45; 187-214.

Appendix

Instructions

1- **General Introduction (given by experimenter)**

There are two activities to be done today. First you will be doing some basic data entry. The second task is to complete a short questionnaire.

For the data entry activity there will be four different periods of 8 minutes each. In each period, periods 1 to 4, you will be asked to enter data for us onto the computer. The data to be entered are patients' laboratory test results. You will be able to earn money for the data entry you do, but you can also decide to use the time in each period to do your own things, such as reading or studying or SMSing your friends.

There are two different types of laboratory reports that you will be given to enter:

- A basic report with only a Full blood count and U&E test
- An expanded report with a FBC and U&E, and liver function test

On your desk you will find an example of each type of report. Each report has a reference number. This is the number you need to enter when asked for the laboratory reference number.

When you move onto data entry, the numbers you type are only entered on the system once you click the SUBMIT DATA button at the bottom. You can click also to submit a partially completed form. **[SLIDE]** In this example, it is indicated that some test results have already been entered on the system by somebody else. Therefore, it is not necessary for you to enter them again.

To make sure everyone has understood the data entry task, we will have a short training period where you can enter the 2 specimen forms on the desk in front of you.
Now to explain the data entry task. As I said there will be 4 different periods. You have four envelopes in front of you, labelled period 1 to period 4. In each envelope there are 15 forms provided for entry during that period. You can only enter the forms allocated for a particular period. The forms should be in numerical sequence at present but you can enter them in any order you like. The maximum number of forms you can enter in the 8 minutes is 15 but we do not expect you to finish all of them. We ask that you do not talk to each other or share information during these tasks.

For agreeing to take part in the study, you have already earned R50. You can earn additional money for your data entry. When you start the activity, a screen will provide some instruction relevant for each period. You will have 90 seconds to read the instructions.
How much you will make depends on what you do in the different data entry periods and also partly on chance. Additional earnings for data entry will be paid in cash at the end of the session. However, we will not pay you for all 4 periods. We will only pay for one of the periods. The period which we will pay will be selected by somebody in the class throwing a die at the end to pick a number between 1 and 4.

2- <u>Screen Instructions</u>

'NO PATIENT BENEFIT' treatment

- **FFS**

You are now going to enter the laboratory reports contained in the envelope labelled "PERIOD 1". In this period, you will be paid R1 for each number you enter.
You will be paid R1for each number you enter irrespective of whether or not it is correct, and irrespective of whether or not it has already been entered on the system.
For example, if "PERIOD 1" is drawn for payment in the random lottery, and if during this period you enter a total of 125 numbers, you will earn R1x 125 = R125 (in addition to the R50 participation fee).

- **Capitation**

You are now going to enter the laboratory reports contained in the envelope labelled "PERIOD 1".
In this period, you will be paid R15 for each basic patient report (FBC and U&E) and R20 for each expanded report (FBC, U&E and LFT). You will be paid for the number of patient reports you do, irrespective of whether or not the individual entries are correct, and irrespective of whether or not some information has already been entered.
For example, if PERIOD 1 is drawn for payment in the random lottery, and if during the period you do 3 basic reports and 4 expanded reports, you will earn 3 x R15+ 4 x R20= R125 (in addition to the R50 participation fee).

- **SALARY instructions**

You are now going to enter the laboratory reports contained in the envelope labelled "PERIOD 1". In this period, you will be paid R125 for the entire period. This payment does not depend on the number of reports or individual entries you make. You will be paid R125 irrespective of whether or not the individual entries are correct, and irrespective of whether or not some information had already been entered on the system.
For example, if PERIOD 1 is drawn for payment in the random lottery, you will earn R125 (in addition to the R50 participation fee).

'PATIENT BENEFIT' Treatment

- **FFS on-screen instructions**

You are now going to enter the laboratory reports contained in the envelope labelled "PERIOD 1".
In this period, you will be paid R1 for each number you enter. You will be paid R1 for each number you enter irrespective of whether or not it is correct, and irrespective of whether or not it has already been entered on the system.

In addition, for each number entered correctly we will send R0.50 to the charity of your choice to provide care to patients (you will choose the charity from a list on the next screen). The charity will be paid only if the number is entered correctly, and if it has not already been entered on the system.
For example, if "PERIOD 1" is drawn for payment in the random lottery, and if during this period you enter a total of 125 numbers, you will earn R1x 125 = R125 (in addition to the R50 participation fee).

If 100 of the numbers are found to be correct, we will also pay 100 x R0.50 = R50 to the charity of your choice.

- **Capitation on-screen instructions**

You are now going to enter the laboratory reports contained in the envelope labelled "PERIOD 1". In this period, you will be paid R15 for each basic patient report (FBC and U&E) and R20 for each expanded report (FBC, U&E and LFT). You will be paid for the number of patient reports you do, irrespective of whether or not the individual entries are correct, and irrespective of whether or not some information has already been entered.

In addition, for each individual entry correctly made we will send R0.50 to the charity of your choice to provide care to patients (you will choose the charity from a list on the next screen). The charity will be paid only if the individual entry is correct, and if it has not already been entered on the system.

For example, if ”PERIOD 1”is drawn for payment in the random lottery, and if during the period you do 3 basic reports and 4 expanded reports, you will earn 3 x R15+ 4 x R20= R125 (in addition to the R50 participation fee). In addition, if 100 of the individual entries are found to be correct, we will pay 100 x R0.50 = R50 to the charity of your choice.

- **Salary on-screen instructions**

You are now going to enter the laboratory reports contained in the envelope labelled "PERIOD 1". In this period, you will be paid R125 for the entire period. This payment does not depend on the number of reports or individual entries you make. You will be paid R125 irrespective of whether or not the individual entries are correct, and irrespective of whether or not some information had already been entered on the system.

In addition, for each individual entry correctly made we will send R0.50 to the charity of your choice to provide care to patients (you will choose the charity from a list on the next screen). The charity will be paid only if the individual entry is correct, and if it has not already been entered on the system.

For example, if “PERIOD 1”is drawn for payment in the random lottery, you will earn R125 (in addition to the R50 participation fee). In addition, if you enter 100 numbers correctly, we will pay 100 x R0.50 = R50 to the charity of your choice.

Fig. A1: Example of a laboratory form

LABORATORY REPORT

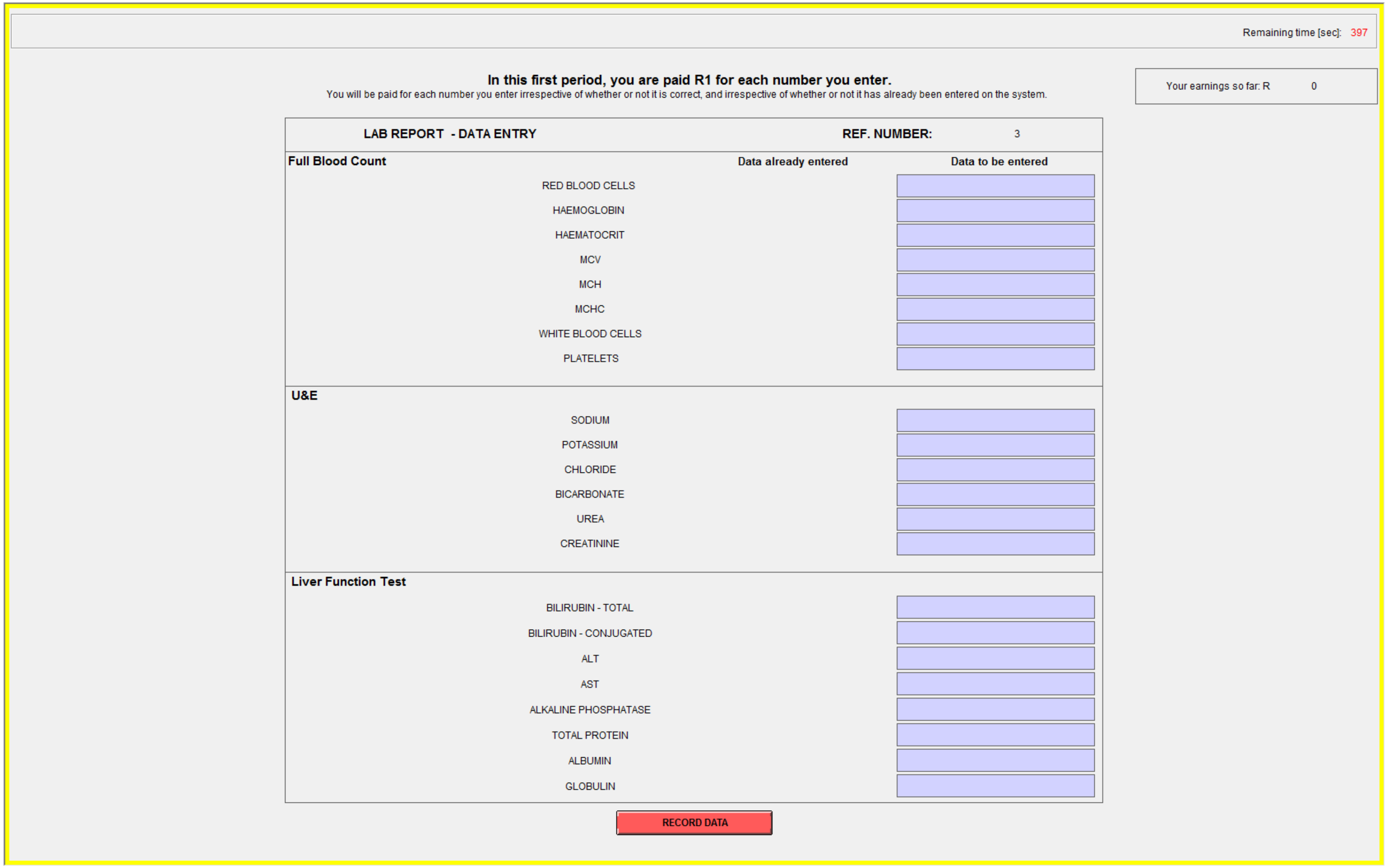
REFERENCE NUMBER 421

Patient age: 29

HAEMATOLOGY AND BIOCHEMISTRY RESULTS

Test		Result	Units	Reference Range
Full Blood Count				
	RED BLOOD CELLS	**5.8**	x 10^{12}/L	4.5 - 6.5
	HAEMOGLOBIN	**15.2**	g/dL	13.8 – 18.8
	HAEMATOCRIT	**47.2**	%	40 - 56
	MCV	**89.8**	fL	79 - 100
	MCH	**27.5**	pg	27 - 35
	MCHC	**35.1**	g/dL	29 - 37
	WHITE BLOOD CELLS	**7.2**	x 10^{9}/L	4.0 – 12.0
	PLATELETS	**317**	x 10^{9}/L	150 - 450
U&E				
	SODIUM	**142.5**	mmol/L	135 - 150
	POTASSIUM	**3.7**	mmol/L	3.5 - 5.1
	CHLORIDE	**103.2**	mmol/L	98 - 107
	BICARBONATE	**23.5**	mmol/L	21 - 29
	UREA	**6.5**	mmol/L	2.1 - 7.1
	CREATININE	**95.8**	µmol/L	80 - 115
Liver Function Test				
	BILIRUBIN - TOTAL	**6.2**	µmol/L	2 - 26
	BILIRUBIN - CONJUGATED	**6.0**	µmol/L	1 - 7
	ALT	**10.9**	IU/L	0 - 40
	AST	**16.5**	IU/L	15 - 40
	ALKALINE PHOSPHATASE	**127.6**	IU/L	53 - 128
	TOTAL PROTEIN	**66.1**	g/L	60 - 80
	ALBUMIN	**35.2**	g/L	35 - 50
	GLOBULIN	**21.7**	g/L	19 - 35

Fig. A2: screen capture of a data entry mask for an expanded form



Remaining time [sec]: 397

Your earnings so far: R 0

In this first period, you are paid R1 for each number you enter.

You will be paid for each number you enter irrespective of whether or not it is correct, and irrespective of whether or not it has already been entered on the system.

LAB REPORT - DATA ENTRY	REF. NUMBER:	3
Full Blood Count	**Data already entered**	**Data to be entered**
RED BLOOD CELLS		
HAEMOGLOBIN		
HAEMATOCRIT		
MCV		
MCH		
MCHC		
WHITE BLOOD CELLS		
PLATELETS		
U&E		
SODIUM		
POTASSIUM		
CHLORIDE		
BICARBONATE		
UREA		
CREATININE		
Liver Function Test		
BILIRUBIN - TOTAL		
BILIRUBIN - CONJUGATED		
ALT		
AST		
ALKALINE PHOSPHATASE		
TOTAL PROTEIN		
ALBUMIN		
GLOBULIN		

RECORD DATA

Fig. A3: screen capture of a data entry mask for a basic form

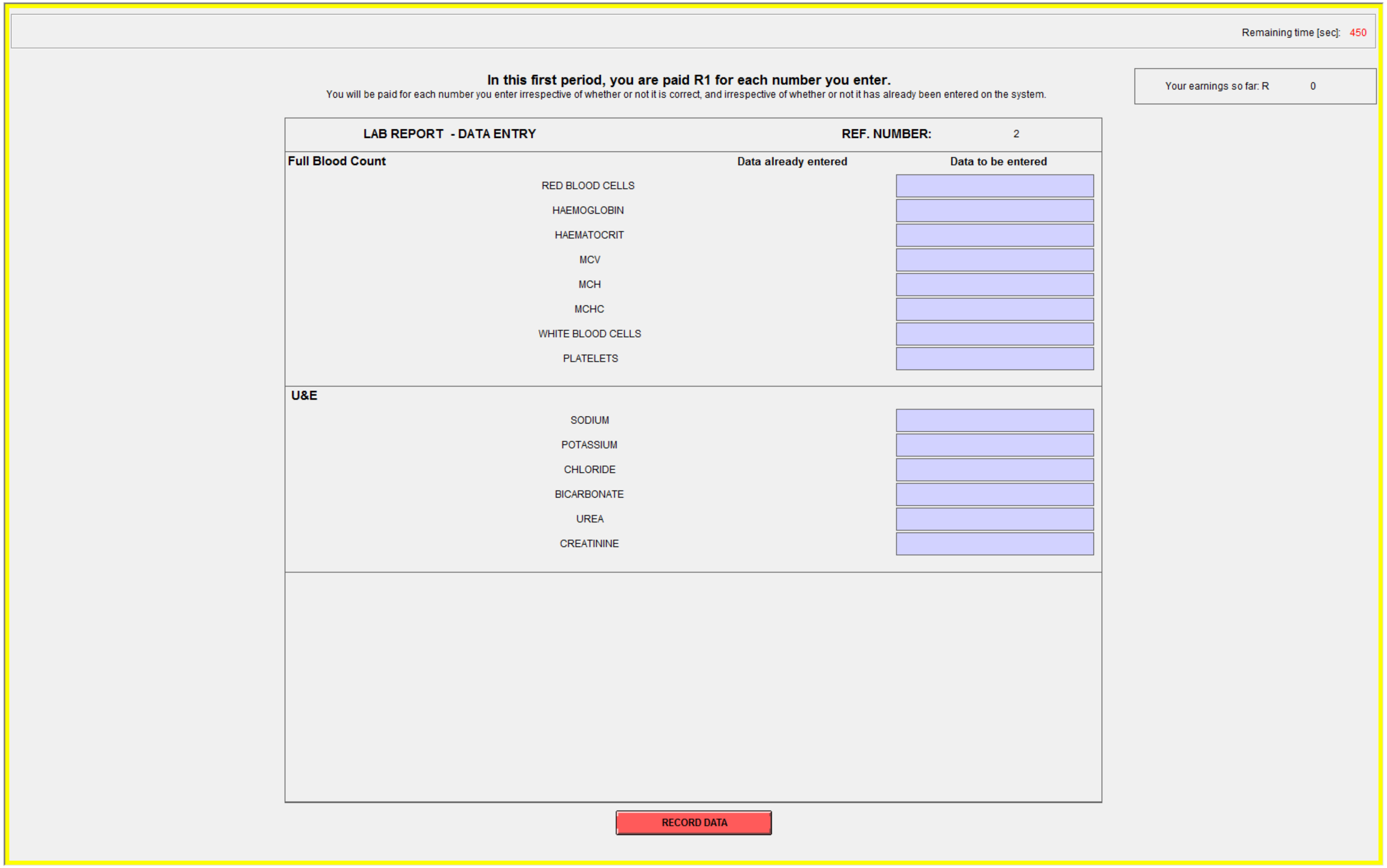


Fig. A4: Screen capture of a data entry mask with information loaded on the system

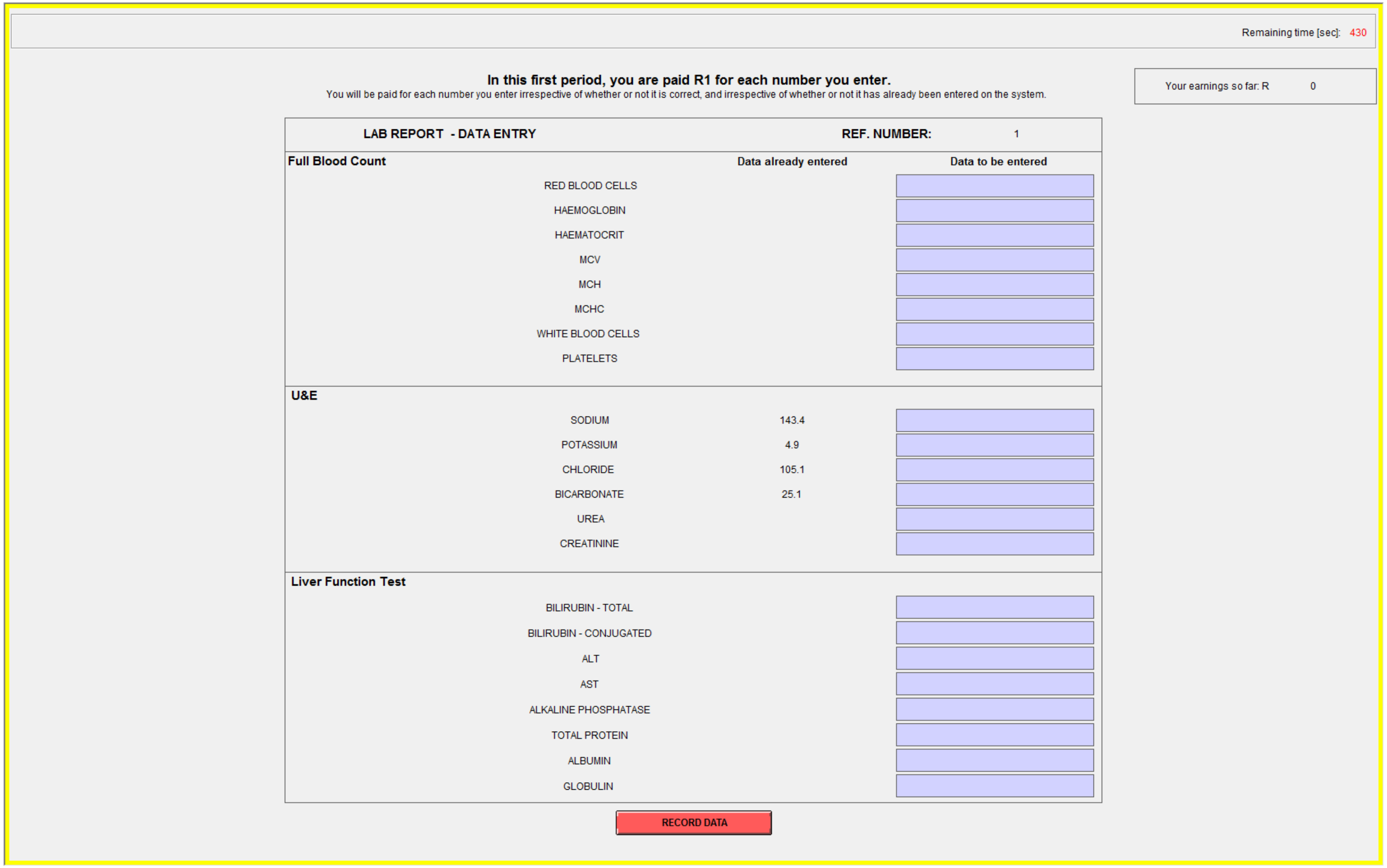


Figure A5: Quantity-quality trade-off under the three different remuneration schemes in the two treatments

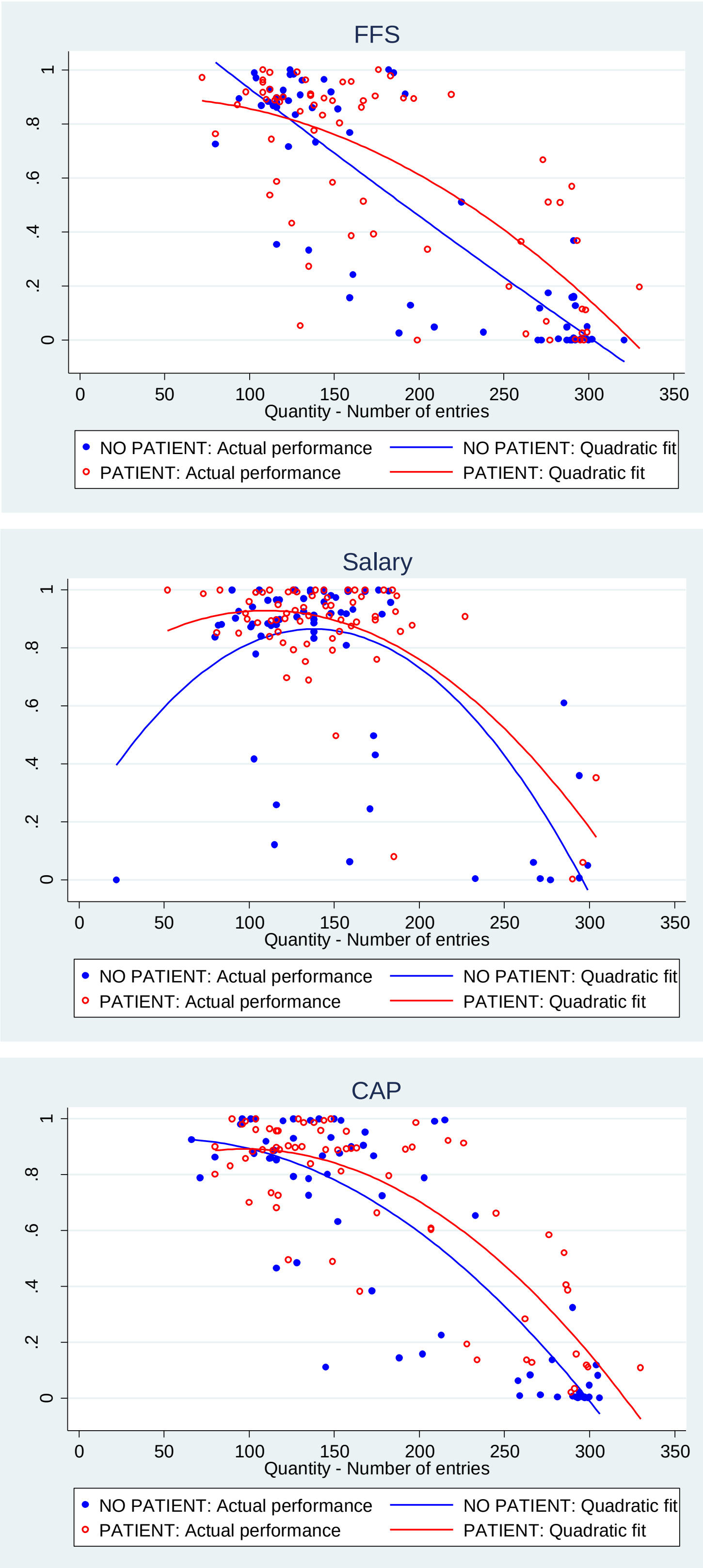


TABLE A1: RANDOM-EFFECT LOGIT REGRESSIONS OF THE IMPACT OF REMUNERATION SCHEMES ON CREAM-SKIMMING, POOLED DATA

	1= put shorter forms first		
	(1)	(2)	(3)
SAL	-0.330	-0.684	-0.778
	(0.476)	(0.637)	(0.658)
CAP	0.017	-0.216	-0.234
	(0.447)	(0.601)	(0.612)
PATIENT	-0.613		
	(0.586)		
PATIENT*FFS		-1.051	-1.096
		(0.801)	(0.821)
PATIENT *SAL		-0.286	-0.293
		(0.775)	(0.788)
PATIENT *CAP		-0.533	-0.573
		(0.773)	(0.789)
Period 2			0.561
			(0.483)
Period 3			0.697
			(0.473)
Entries made	-0.009**	-0.009**	-0.011**
	(0.004)	(0.004)	(0.004)
Constant	-1.117	-0.889	-1.012
	(0.858)	(0.897)	(0.918)
Observations	396	396	396
Log-Likelihood	-144.1	-143.7	-142.5

Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1