Payment Systems in the Healthcare Industry: An Experimental Study
Of Physician Incentives

By
Ellen P. Green
The views expressed in the Working Paper Series are those of the author(s) and do not necessarily reflect those of the Department of Economics or of the University of Delaware. Working Papers have not undergone any formal review and approval and are circulated for discussion purposes only and should not be quoted without permission. Your comments and suggestions are welcome and should be directed to the corresponding author. Copyright belongs to the author(s).
Payment Systems in the Healthcare Industry: An Experimental Study of Physician Incentives

By Ellen P. Green

Abstract:

Policy makers and the healthcare industry have proposed changes to physician payment structures as a way to improve the quality of health care and reduce costs. Several of these proposals require healthcare providers to employ a value-based purchasing program (also known as pay-for-performance [P4P]). However, the way in which existing payment structures impact physician behavior is unclear and, therefore, predicting how well P4P will perform is difficult. To understand the impact physician payment structures have on physician behavior, I approximate the physician-patient relationship in a real-effort laboratory experiment. I study several prominent physician payment structures including fee-for-service, capitation, salary, and P4P. I find that physicians are intrinsically motivated to provide high quality care and relying exclusively on extrinsic incentives to motivate physicians is detrimental to the quality of care and costly for the healthcare industry.

Keywords: Physician Payment System, Laboratory Experiment, Incentives, Fee-for-Service, Capitation, Salary, Report Cards, Pay-for-Performance, Crowd Out

JEL Classification: C9, I1

 Virginia Tech, Department of Economics, 3016 Pamplin Hall, Blacksburg, VA 24061, United States epgreen@udel.edu. Permanent Address: University of Delaware, Department of Economics, 416C Purnell Hall, Newark, DE 19716, United States. I would especially like to thank my advisor, Sheryl Ball and colleague Adrienne Lucas for their useful comments and suggestions. I am also grateful to Landis Atkinson, Kristen Capogrossi, Ben Congleton, Pamela Cubberly, Ryan Donnelly, Saul Hoffman, Russell Osborne, Roger Congleton and seminar participants at the American Economic Association annual conference, Southern Economic Association annual meeting, Society for Medical Decision Making annual meeting, and International Economic Science Association conference for helpful comments. Support from the Institute for Advanced Study and the Making A Difference Scholarship at Virginia Tech is gratefully acknowledged. All errors are my own.

1. Introduction

Traditionally, physicians are paid for each service that they perform (i.e., fee-for-service). However, in an attempt to curb unnecessary care in the early 1990s, some primary care practices moved away from reimbursing physicians through simple fee-for-service programs to
compensation through capitated rates, or a lump sum per patient for a specific illness. Despite this effort, the cost of health care continued to increase by alarming rates through the 2000s. In the early 2000s, value-based purchasing programs were proposed as a way to encourage better care at lower costs. In 2010 applications of these proposals came to fruition under the Patient Protection and Affordable Care Act.

Value-based purchasing programs (also known as pay-for-performance [P4P]) pay agents for performance according to a set of defined quality measures (e.g., a practice may be compensated based on keeping a percentage of patients with diabetes under a predetermined level of LDL-C or keeping their diabetes “under control”). While P4P payment structures are effective in theory, previous studies of P4P payment structures have shown that P4P failed to improve agent performance and reduce costs (Gillam et al., 2012; Rosenthal et al., 2006; Werner et al., 2011; Werner et al., 2013). Furthermore, the way in which traditional payment structures impact physician behavior is unclear (Gosden et al., 2001; Sørensen and Grytten, 2003). Therefore, the way in which new payment structures will impact physician behavior remains unknown. This paper uses a real-effort laboratory experiment to examine the impact of physician payment structures on behavior.

While physicians often maintain that their behaviors are independent of payment and driven by professional standards of care, many economic theorists argue that payment structure influences a physician’s behavior. For example, some economic theorists argue that, under a capitated payment system, physicians may undertreat patients while treating as many patients as feasible (Matsaganis and Glennerster, 1994; Robinson, 2001; Robinson et al., 2004). While the fee-for-service payment system encourages physicians to over treat patients by paying physicians per service provided.

Empirical research on this topic is mixed. Some studies found that this simple theory accurately predicts the relationship between physician behavior and payment system (Gaynor and Gertler, 1995; Hennig-Schmidt et al., 2011), but other studies have shown that regardless of incentive system, physicians do not change their behavior (Croxson et al., 2001; Devlin and Sarma, 2008). This inconsistency is likely because the studies are set in active physician practices where not all factors can be controlled, accurately measured, or may be endogenously related (Gosden et al., 2000; Gosden et al., 2001). For instance, payment system changes are typically implemented suddenly, in which case a robust comparison of them before and after the modification is not possible. Furthermore, determining if other components of the physician’s payment structure changed simultaneously is difficult (e.g., unaccounted for bonus incentives or converting the internal record keeping system, i.e., electronic medical records). Also, patients
may self-select into practices with specific payment mechanisms, creating another potential source of bias. Lastly, the data required to accurately measure the change to a patient’s health status are not available and researchers are often forced to rely on the self-reported health (Gaynor and Gertler, 1995), which likely differs from reality (Camerer and Hogarth, 1999). For these reasons, in this paper I use a laboratory experiment to isolate the impact of payment systems on physician behavior in a controlled environment.

The experimental literature on physician payment structures is limited. While Fuchs (2000) was the first to acknowledge the potential insights from laboratory experiments, Hennig-Schmidt et al. (2011) were the first to conduct a laboratory experiment with specific applications in health care. Hennig-Schmidt et al. (2011) focused their study on how fee-for-services and capitation influence physician behavior, their main finding was that physician behavior was influenced by payment structures consistent with the predictions of simple theory (i.e., fee-for-service overprovided and capitation underprovided services).

To improve our understanding of the effect of physician payment structures on health care provision, this paper expands on the previous literature through the use a real-effort laboratory experiment to study several prominent payment structures’ impact on behavior in a controlled environment. I test five separate payment structures: fee-for-service (FFS), capitation (CAP), salary, P4P, and report cards. FFS and P4P pay retrospectively, and the others pay prospectively. The experimental design for the present study was derived from the multiple principal agent model (Ellis and McGuire, 1986, 1990) and experiments by (Hennig-Schmidt et al., 2011).

The experiment was designed to imitate the relationships among the physician, patient, and healthcare provider (the physician’s employer). As implemented, student subjects who represented physicians were hired to provide proofreading services for a different group of student subjects who represented their patients. The physicians had to make a decision on how many services to provide and for which patients to provide services. The physicians were remunerated for their services by one of the payment structures mentioned above. The experimental design allowed physicians the opportunity to under or overtreat their patients. The physicians’ actions had direct consequences on their patients’ payoff. All experimental parameters remained constant across treatments, aside from the physician’s payment structure.

I find that intrinsic motivations play an important role in physician decision-making and that retrospective payment structures “crowd out” these motivations. Specifically, when pay was retrospective (i.e., the FFS and FFS with P4P), physicians provided a lower overall quality of services. Conversely, when payment structures paid prospectively (i.e., the salary and CAP),
physicians provided a higher overall quality of service. This disconnect between physician behavior and monetary incentives demonstrates that monetary incentives can “crowd out” intrinsic motivations that would otherwise motivate subjects to complete the task at higher levels of performance.\(^1\)

Furthermore, I find that the retrospective payment structures failed to improve quality of care. Specifically, the physicians contracted by salary provided unnecessary services only 8.5% of the time in the experiment, while the agents contracted by FFS with P4P provided unnecessary services 72.1% of the time. This result is not surprising, as the retrospective payment structures do little to curb unnecessary, wasteful, or harmful services, which are not only costly to the patient (some estimate waste accounts for 33% of total health-care expenditures [Berwick and Hackbarth, 2012]) but are also are oftentimes harmful to the patient’s health. As opposed to dissuading the provision of unnecessary services, P4P systems encourage the overprovision of services listed under their defined quality measures (Petersen et al., 2006).

Based on actual per capita spending on health care in the United States (OECD, 2012), the experimental results suggest that physicians paid under FFS with the P4P payment system would annually provide to each patient nearly $6,000 of unnecessary or harmful services and only about $2,300 of beneficial services.\(^2\) In comparison, measured in the same way, physicians paid under the salary payment system are predicted to provide less than $700 of harmful or unnecessary services and more than $7,000 of necessary services. These numbers, while only suggestive, provide a rough indication of the money wasted based on one of the predominantly used payment systems. However, the approximations do demonstrate the potential benefits that could be gained in the health-care industry by simplifying physicians’ contracts to optimize intrinsic motivations and, therefore, move away from P4P-type payment structures.

2. Data and Methodology: The Dual-Principal Agent Game

The dual-principal agent relationship in the healthcare industry involves two principals interacting with a single agent. The upstream principal (the physician’s employer) in charge of contract design does not receive the services provided by the agent and the downstream principal

---

1 In this application, several intrinsic motivations are potentially at play, for example, self-recognition (Deci and Ryan, 1985), confidence management (Bénabou and Tirole, 2003), other regarding behavior (Hoffman et al., 1994); for a review of intrinsic motivations see Deci et al. (1999).
2 This number is an approximation based on the average total spending per capita in the United States. Unfortunately, this number aggregates spending from all payment structures. Therefore, because heterogeneity in payment systems is already in place, the waste stated here might over state the magnitude of the problem.
(the patient) receiving the services, does not directly pay the agent (the physician). In this paper, I replicate this relationship in a real-effort laboratory experiment.

In many respects, the Dual-Principal Agent (DPA) game utilizes the same methodology as principal-agent work by Fehr and Schmidt (2004) and Hennig-Schmidt et al. (2011). In the DPA game, the players representing the health-care provider, the patient, and the physician interact in a one-time application of their contract. The single interaction eliminates the influences of physician reputation, promotions, and/or the possibility of contract termination on the outcome of the contract. The patients have no authority over with which physician they were paired, preventing patients from employing screening systems in their efforts to improve outcomes. Physicians are not given the opportunity to select their upstream principals (employers) based on the payment system offered, eliminating the possibility of adverse selection. By eliminating the opportunity for self-selection into a payment system, the DPA game focused exclusively on the effects of the different payment systems on agent behavior, something that cannot readily be achieved in a non-laboratory setting.

The major distinction between the DPA game and above-mentioned principal-agent experiments is that the DPA game utilizes a real-effort task, rather than a chosen effort task. In Fehr and Schmidt (2004) and Hennig-Schmidt et al. (2011), agents’ earnings were a function of the cost for the “work product” generated and their payment system. In those studies, the value of work produced and the cost of effort were numbers from a table that the subjects selected, requiring no real effort to be exerted (i.e., a chosen effort task). In the current DPA game, an agent selects her own effort in a real-effort task, which is used to determine her downstream principal-counterpart’s final payment and her own final payment. For simplicity, the remainder of the paper refers to employers (upstream principals), physicians (agents), and patients (downstream principals), but the findings are applicable to any similarly situated agent relationship.
2.1 Experimental Design

Each participant in the DPA game represented a specific role in the dual-principal agent relationship. The experimenters functioned as the health-care employers, as they assigned the payment systems and provided payment to the physicians. Each subject acted as either a physician or a patient in the relationship.³

The DPA game took place in two phases. In Phase I, 10 subjects were recruited to represent patients. Each patient was endowed with $25.00 and asked to proofread 10 essays. Each essay had a total of 10 typographical or spelling errors that the patients were asked to correct. For each error that the patient failed to correct, a penalty of $0.25 was taken from his endowment.⁴ The number of errors that remained represented the ideal number of services that the patient should receive from their physician.

In Phase II of the game, new subjects were recruited to represent the physicians. The physicians provided proofreading assistance to the 10 patients from Phase I of the game. The physicians’ task was to correct the errors that were missed by the patients in Phase I. Each physician was eligible to receive one edited essay from each patient. Therefore, the Phase II physicians could assist up to 10 distinct patients, if they wished.

The passages presented to the physicians differed from those submitted by the patients on two margins. First, to assist the physicians in the proofreading task, missed errors were

Example 1:
The root to the stop was familiar from countless hours of riding bikes, roller-blading, and the like, but today, it seemed completely foreign to me.

³ I assume there was no other-regarding behavior toward the experimenter based on the results of Frank, B., 1998. Good news for experimenters: subjects do not care about your welfare. Economics Letters, 61(2), 171-174.
⁴ Failure to identify any of the mistakes resulted in a payment of 0, ($25-10*$0.25*10=0)
highlighted in turquoise. This assistance was provided to reflect the knowledge of a practicing physician, who would be more capable of determining an ailment than the patient himself (i.e., would be more capable of finding an error in a proofreading task). Second, for each error that was missed by a patient in Phase I of the game, the physician was shown two additional turquoise-highlighted words within the essay. Example 1 provides a sample sentence as seen by the physician. The patient failed to note one error that should have been corrected (“root” should have been changed to “route”). The other two highlighted words, “was” and “foreign,” were not errors to be corrected but were highlighted to permit the overprovision of services to patients. Each change by a physician to a highlighted word counted as a service provided and was remunerated accordingly, even the changes that were incorrectly made.5

The treatment variable in the DPA game was the payment structure. The treatments were payment through FFS, CAP, salary, CAP with report card, CAP with P4P, and FFS with P4P. FFS physicians were paid a set rate for each “error” they corrected, whether or not their modification of the text was actually a correction. CAP physicians were paid a set rate for each patient to whom they provided at least one service. Salary physicians were paid a set rate for participating in the experiment (refer to Table 1 for rates). The CAP with report card payment system employed the CAP payment system with the added condition that, every two minutes during the proofreading task, a table with the number of changes currently made to the text by the other physicians in the session was available on the screen, effectively providing physicians with information about how many services other physicians were “prescribing” during the session.

In CAP with P4P and FFS with P4P, the physicians earned a bonus if they were able to meet a predetermined quota of errors designed to replicate quality measures. Under the P4P

---

5 In the example provided, a change of “was” to “were” would have received payment even though the change was incorrect. Words that were not highlighted could not be edited.
payment structure, the “errors” that satisfied the quota requirement were uniquely identified in the text through yellow highlighting. There were 27 yellow errors in total. Only 9 of the yellow highlighted sections required editing. The 18 other yellow “errors” were highlighted to permit overtreatment. The physicians that were paid under FFS with P4P and CAP with P4P systems earned a bonus of $1.00 if they edited at least 9 out of 27 yellow errors. The yellow errors mirror the treatments in the health-care industry under the P4P payment structures, such as mammograms or colonoscopies, for which physicians are given a bonus if they provide a specific percentage of their patients with the service (Armour et al., 2004). The physicians did not earn the bonus for correcting any predetermined quantity of normal errors (i.e., errors highlighted in turquoise). The P4P payment was additional to the base payment system, i.e., FFS or CAP.

In each session of the experiment, all physicians were assigned to the same payment system as determined by their employer (the experimenter) before the experimental session. The physicians were unaware that other sessions were paid differently. At the end of their task, the physicians were paid for the changes made according to the conditions of the physician’s particular payment system.

The physician’s behavior impacted the patient’s payoff, as it would have in a medical setting. For each patient error that a physician correctly identified, the patient received $0.15 after Phase II was completed. This value reflects the increase in the patient’s utility from the beneficial service. The difference between the amount the patient lost in Phase I and the amount she was reimbursed in Phase II from the beneficial service reflects the monetary cost of the service as well as the notion that the patient’s health status is not necessarily restored to 100% of its original value after a service is provided. To reflect the cost to a patient of being overprovided with services, for each instance that a physician incorrectly edited an actual mistake or changed a highlighted word that was not a true mistake, the patient’s earning decreased by $0.05. This demonstrates both the monetary cost and the harm to the patient’s health from unnecessary services. In this way, patients were rewarded for the equivalent of a correct diagnosis by a physician and penalized for being incorrectly prescribed a treatment to reflect the change in utility.

---

6 In total, all treatments had the same number of “errors”, i.e., the P4P payment structures had the same number of total errors (yellow and turquoise) as those without P4P. The “errors” highlighted in yellow under P4P were highlighted in turquoise under the non-P4P payment structures.

7 Changes that were made incorrectly to the yellow errors were still considered services provided and counted toward the physician’s quota.

8 Patients lost $0.25 for each error they missed in Phase I and were only reimbursed $0.15 of this amount if the error was correctly identified by the physician. Therefore, the dominant strategy for Phase I patients was to correct as many errors as possible.
based on both harmful and helpful services provided. The exact values for the patient’s reimbursement scheme come from simulations that were run before the experiment.

2.2 Procedure

The experiments were conducted in the Economics Research Laboratory at Virginia Tech. Students were recruited from Virginia Tech Principles of Economics courses. A total of 136 subjects participated in the experiment. Several sessions were held for both Phase I and Phase II. No subject participated in more than one session.

In Phase I, recruited subjects played the role of the patient in the dual-principal agent relationship. In each session, as the participants entered the laboratory lobby, they completed a consent form and waited for further instruction. The subjects were then escorted into the laboratory’s classroom and asked to edit the 10 essays within 50 minutes. The subjects could not view the other participants’ actions in the task. Subjects were free to leave before the allotted time if they were satisfied with their effort. Once the task was completed, subjects were paid individually according to their success in completing the task. At this time, the subjects were also given further information about how some of their endowment could potentially be earned back through assistance from participants in Phase II of the experiment. If the subjects wished to collect this additional payment, contact information was collected. Between Phase I and Phase II of the experiment, the proofread essays from the subjects in Phase I were entered into a computer database to be used in Phase II.

For Phase II of the game, different subjects were recruited to represent physicians. This phase was conducted over 13 sessions. In each of these sessions the subjects entered the laboratory lobby, completed a consent form, and waited for further instruction. Once all of the subjects were present, they were escorted into the computer laboratory and asked to sit at a terminal. When seated, the subjects were given instructions for their task. The details of the payment system were provided to the subjects both verbally and in text in each session and subjects were given the opportunity to ask questions about their task. After the instructions were complete and all questions had been answered, the physicians were asked to complete a quiz on the details of their payment system to ensure that they fully understood how their actions would affect both their payoff and the patients’ payoffs. The experiment continued only when all subjects were able to successfully answer the quiz questions.

The Phase II subjects then performed the proofreading task. The subjects could start proofreading any of the 10 essays assigned to them and were able to go back and forth between essays as they wished. The computer interface allowed Phase II subjects to determine how many
and for which of the Phase I subjects they would provide services; the physicians could help none or all of the Phase I subjects if they wished.\(^9\) Each physician in Phase II was assigned the same patient panel, that is, the same 10 essays. However, the patients from Phase I would only be reimbursed based on the behavior of one physician from Phase II. The physicians were informed of that the patients would only be reimbursed by one physician in their instructions. The physician that was selected for each patient was determined using a random number generator.

To prevent proofreading out of boredom, the subjects were able to leave the proofreading task whenever they wished. After the subjects were finished with their proofreading tasks, they completed a brief survey with questions drawn from Robinson et al. (1991) that collected data on the subject’s trustworthiness, altruism, reciprocity, and socioeconomic status, as well as how interesting they felt the task was.\(^{10}\) The physicians were then paid individually according to the parameters of the relevant payment system. When Phase II was concluded, the total reimbursement for participants in Phase I was calculated, and the participants were contacted to receive their further compensation.

### Table 2: Physician's Services by Treatment Reported as Averages

<table>
<thead>
<tr>
<th>Payment Mechanisms</th>
<th>Retrospective</th>
<th>Prospective</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FFS</td>
<td>FFS with P4P</td>
</tr>
<tr>
<td>Harmful</td>
<td>20.32 (32.66)</td>
<td>69.05 (50.30)</td>
</tr>
<tr>
<td>Beneficial</td>
<td>33.00 (10.80)</td>
<td>26.71 (10.95)</td>
</tr>
<tr>
<td>Total</td>
<td>53.31 (34.37)</td>
<td>95.76 (46.48)</td>
</tr>
</tbody>
</table>

Note: Standard deviation appears in parenthesis under sample means. *Harmful services*: the number of incorrect amendments to the text. *Beneficial services*: the number of correct amendments to the text. *Total services*: the sum of beneficial services and harmful services.

### 3. Results

#### 3.1 Physician Behavior

Table 2 provides summary statistics by payment structure of the physicians’ actions in the DPA game. Actions are disaggregated into harmful, beneficial, and total services. *Harmful*

---

\(^9\) Physicians did not know the identity of the patients, just their anonymized patient number.

\(^{10}\) The altruism, trustworthiness, reciprocity, socioeconomic, task-interest data collected were used in regression models to test their correlation with agent behavior. None of these variables had a statistically significant correlation with agent behavior.
services are the number of incorrect changes the physician made to the essays. Beneficial services are the number of correct amendments to the text the physician made. Total services are the sum of beneficial services and harmful services.

As expected, physicians paid by the retrospective payment systems (FFS and FFS with P4P) provided more total services when compared with the other four prospective payment systems (salary, CAP, CAP with P4P and CAP with report card). This is evident in Figure 1. Specifically, physicians compensated through the FFS with P4P provided more than two and a half times as many services (95.76) as the salaried (36.13), CAP with report card (31.15), CAP with P4P (33.47), or CAP (36.14) compensated physicians. The physicians compensated through FFS provided about one and a half times as many services (53.31) as the salaried, CAP with report card, CAP with P4P, or CAP physicians on average. To test the significance of these results, I used the two-sample Mann-Whitney U test based on ranks to test whether the amount of total services came from the same distribution. The results are reported in Appendix: Table 2a. As expected, the differences in total services provided amongst payment structures that were large in magnitude and those that were small in magnitude were statistically different.

Across all payment structures, the number of beneficial services delivered varied from a low of 26.71 provided by physicians under the FFS with P4P scheme to a high of 33.05 provided by the salaried physicians. Figure 2a reports, by payment structure, the average number of
beneficial services. Again, I used the two-sample Mann-Whitney U test based on ranks to compare distributions. A table of Mann Whitney U P-values comparing the beneficial services provided by payment structure appears in Appendix 1: Table 2b.

Although these differences are smaller than for total services (Figure 1), the difference between the average number of beneficial services provided under CAP (32.74) and salary (33.05) payment systems are significantly different than those provided under FFS with P4P (26.71) and CAP with report card (27.60). Even though physicians under CAP and salary did not have a monetary incentive that linked behavior with payment, the physicians provided a greater number of beneficial services than the retrospective payment structures, which suggests that physicians had intrinsic motivations for providing beneficial services to the patient. This conflicts with classical literature in the economics field that predicts the physicians compensated through salary or CAP would underprovide services (Hart and Holmstrom, 1987; Holmstrom and Milgrom, 1991). Instead, this finding is consistent with previous literature in psychology as well as some new experimental literature in economics that demonstrated that the monetary incentives of FFS and FFS with bonus potentially muddle or crowd out the positive influences of intrinsic motivations (Ariely et al., 2009; Ballou, 2001; Boudreau et al., 2012; Bénabou and Tirole, 2003; Camerer et al., 1997; Deci, 1971; Fehr and Schmidt, 2004; Holmstrom and Milgrom, 1991).

The variation in total services (Figure 1) provided by the physicians comes primarily from the number of harmful services provided under the different payment structures. The large variation in harmful services is evident in Figure 2b, which shows the average number of harmful services by payment structure. The average number of harmful services ranges from 69.05 under the FSS with P4P payment scheme to 3.09 under the salary payment scheme.

Physicians compensated using FFS with P4P provided 69.05 harmful services, significantly more than physicians under any other payment structure. The physicians who conducted the second highest number of harmful services were those paid under FFS (20.32). These physicians provided a significantly greater number of harmful services to the patient than physicians paid via the salary (3.09), CAP with report card (3.55), CAP (3.39), and CAP with P4P (4.24) payment systems. This suggests that paying physicians retrospectively on services provided may be detrimental to quality of care, which I explore further under the next section.

---

11 A table of Mann Whitney U P-values comparing the harmful services provided by payment structure is located in Appendix 1: Table 2c.
Next, I explore the impact of the payment structures on the patients. Table 3 summarizes the monetary impact of physician behavior on patients, separately by payment structure. Negative impact and positive impact are linear transformations of the harmful and beneficial services described above. Negative impact is the cost (-$0.05 per error) the patient incurred due to an incorrectly identified error (harmful services) by the physician. Positive impact is the benefits (+$0.15 per error) that the patient received from correctly identifying errors (beneficial services) by the physician. Net benefit is the difference between the positive impact and negative impact, measuring the direct effect the physician’s actions had on the patient’s income, a proxy for the quality of care.

The salary and CAP compensation schemes resulted in the most positive changes in income: patient payments averaged of $4.80 and $4.74 respectively. Both of the FFS payment structures—FFS ($3.93) and FFS with P4P ($0.55)—continued to underperform relative to all of the other payment structures. This result implies that paying physicians retrospectively on services provided may be less beneficial to the patients’ outcome. A two-sample Mann-Whitney U test based on ranks showed that the CAP and salaried physician continued to outperform other payment structures and had a significantly greater net benefit than FFS with P4P ($0.55) and CAP with report card ($3.96). A table of the two-sample Mann-Whitney U test based on ranks P-values comparing the net impact on the patient by payment structure is located in Appendix 1: Table 3a.
3.3 Impact on Employer

Table 4 summarizes the impact on the physician’s employer by costs, net benefit, and net benefit-to-cost separated by payment structure. *Cost* is the average payment made to the physician in the experiment. *Net benefit* is calculated in the same manner as impact on patient as described in section B. The *net benefit-to-cost ratio* is the average net benefit divided by the cost of employing the physician.

The highest cost to the employer was the salary payment system ($25.00) followed closely by the CAP payment system ($24.35).\(^{12}\) Somewhat surprisingly, I did not find that the FFS and FFS with P4P payment structures resulted in the highest cost, even though their incentives created a positive relationship between income (cost to the employer) and the number of services provided. This is most likely an artifact of the size of the salary relative to FFS, rather than a change in behavior caused by the payment structures or a consequence of the time limit on the game.\(^ {13}\)

Finally, I explore costs in relation to the net benefits of the services provided. Row 3 of Table 4, reports the net benefit-to-cost ratio. The results are somewhat surprising: the FFS payment structure resulted in the highest net benefit-to-cost ratio. For each dollar spent on the physician, the benefit was higher on average when physicians were paid via FFS ($0.49) when compared with the other payment structures. Conversely, FFS with P4P ($0.13) continued to

<table>
<thead>
<tr>
<th>Retrospective</th>
<th>Payment Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFS</td>
<td>FFS with P4P</td>
</tr>
<tr>
<td>Costs</td>
<td>$10.76</td>
</tr>
<tr>
<td>(6.95)</td>
<td>($9.55)</td>
</tr>
<tr>
<td>Net Benefit</td>
<td>$3.93</td>
</tr>
<tr>
<td>(2.31)</td>
<td>(3.56)</td>
</tr>
<tr>
<td>Net Benefits/Cost</td>
<td>$0.49</td>
</tr>
<tr>
<td>(0.25)</td>
<td>(0.27)</td>
</tr>
</tbody>
</table>

Note: Standard Deviations appear parentheses under the sample mean. *Cost*: the average payment made to the physician. *Net benefit*: impact on patient from section B. *Net benefit-to-cost ratio*: calculated as the average net benefit divided by the cost of employing the physician.

---

\(^{12}\) A table of Mann Whitney U P–values comparing the costs to employer by payment structure is located in Appendix 1: Table 4a.

\(^{13}\) Some subjects did change almost all of the errors available to them in the given time frame. Therefore, the time limit did not appear to be a binding constraint.
underperform and had a significantly lower net benefit-to-cost.\textsuperscript{14} Although the differences are not statistically significant, the results indicate the following ranking from lowest to highest net benefit-to-cost: FFS ($0.13), salary ($0.19), CAP ($0.19), CAP with report card ($0.17), CAP with P4P ($0.18), and FFS with P4P ($0.49).

4. Discussion

While our laboratory experiment provides additional insights in the decision-making of agents within the healthcare industry, some of the simplifications raise concerns. First, students proof-reading an essay are different than physicians making medical decisions. While students in the DPA game directly impacted their counterpart, the impact was not to the same extent as a physician’s impact on their patient—and I do not try to argue that here. I argue that the results indicate that payment structures influence agent behavior and not necessarily in the manner that policy makers expect. Additionally, although many physicians argue that their decision-making is not based on monetary reward, overtreatment does exist (Korenstein et al., 2012) and monetary rewards do influence physician behavior in the ways we see here (Paxton et al., 2012).

Second, the healthcare industry has cost layers that are not represented in the DPA game. However, the goal here was not to replicate precisely the healthcare industry, but to simplify reality to isolate the impacts of various payment systems on agent behavior. By building onto the DPA game in future experiments, the motivations behind physician decision-making and the effects of each changing variable can be better understood. Starting with the fundamentals, as studied in this experiment, is important to observe the impact of each of the influencing variables.

5. Conclusion

This paper explored how six prominent physician payment systems (FFS, FFS with P4P, CAP, CAP with P4P, salary, and CAP with report card) influenced physician behavior in the Dual-Principal Agent (DPA) game. The DPA game was configured to approximate the relationship between a physician, his patient, and the health-care provider (the physician’s employer). The DPA game simplified this relationship to isolate the impact that the six payment systems had on physician behavior without the other influences that would occur in settings in which payment structure might be endogenously determined or be part of a larger change in practices. For example, payment structure could endogenously change along with or lead to

\textsuperscript{14} A table of Mann Whitney U P-values comparing the net benefit-cost ratio by payment structure is located in Appendix 1: Table 4b.
additional changes in demographics, organizational hierarchy, context-specific decisions, reimbursement, variables that were held constant in the laboratory setting.

I find that the payment structures that remunerated physicians retrospectively by their behavior crowded out intrinsic motivations that would otherwise improve the outcome for the patient. Specifically, when differentiated by quality of services provided, payment systems with retrospective reimbursements (i.e., FFS, and FFS with P4P) resulted in the lowest overall quality of services to the on patient. In contrast, the payment systems with prospective reimbursements (i.e., salary, CAP, CAP with report card and CAP with P4P) resulted in a higher overall quality of services.

The crowding out effect was most apparent in the FFS and FFS with P4P payment systems. The FFS system paid physicians on a per-service basis, placing more focus on the maximization of income relative to the other payment systems. For instance, physicians paid via salary were able focus on their task without concern for their income, because it was guaranteed regardless of their behavior. Therefore, these physicians’ decision making was determined by their intrinsic motivation to help the patient. However, FFS physicians maximized their income by increasing the number of services provided (both beneficial and harmful) and, thereby, hurt the patients. The intrinsic motivations that lead to high-quality of services are overridden by the “hidden costs” of the payment structures. Agents appear to be intrinsically motivated and complex behavior-based payment structures “crowded out” this motivation.

In conclusion, the impact that nonmonetary incentives in the DPA game had on agent behavior supports the use of simple payment systems in multiple-principal agent relationships such as the health-care industry. The results demonstrated that agent behavior is impacted by payment systems but not as straightforwardly as classic theory predicts. In some cases, intrinsic motivators were able to improve the quality of services provided by the physicians when the payment incentives did not crowd them out. However, more research needs to be conducted to understand physician behavior before implementing large-scale changes to the health-care industry’s payment system such as the value-based purchasing program.
References:


### Appendix 1: P-Values for all Two-Sample Mann Whitney U-tests

#### Table 2a: P-values from Mann Whitney U-Test on Agent's Actions for Total Services

<table>
<thead>
<tr>
<th>Payment Mechanisms</th>
<th>CAP</th>
<th>CAP with P4P</th>
<th>FFS</th>
<th>FFS with P4P</th>
<th>Report Card</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP</td>
<td>-</td>
<td>0.59</td>
<td>0.08</td>
<td>0.00</td>
<td>0.03</td>
<td>0.66</td>
</tr>
<tr>
<td>CAP with P4P</td>
<td>0.59</td>
<td>-</td>
<td>0.05</td>
<td>0.00</td>
<td>0.65</td>
<td>0.40</td>
</tr>
<tr>
<td>FFS</td>
<td>0.08</td>
<td>0.05</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td>FFS with P4P</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Report Card</td>
<td>0.03</td>
<td>0.65</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
<td>0.07</td>
</tr>
<tr>
<td>Salary</td>
<td>0.66</td>
<td>0.40</td>
<td>0.07</td>
<td>0.00</td>
<td>0.07</td>
<td>-</td>
</tr>
</tbody>
</table>

#### Table 2b: P-values from Mann Whitney U-Test on Agent's Actions for Beneficial Services

<table>
<thead>
<tr>
<th>Payment Mechanisms</th>
<th>CAP</th>
<th>CAP with P4P</th>
<th>FFS</th>
<th>FFS with P4P</th>
<th>Report Card</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP</td>
<td>-</td>
<td>0.40</td>
<td>0.97</td>
<td>0.02</td>
<td>0.02</td>
<td>0.85</td>
</tr>
<tr>
<td>CAP with P4P</td>
<td>0.40</td>
<td>-</td>
<td>0.33</td>
<td>0.55</td>
<td>0.69</td>
<td>0.38</td>
</tr>
<tr>
<td>FFS</td>
<td>0.97</td>
<td>0.33</td>
<td>-</td>
<td>0.05</td>
<td>0.07</td>
<td>0.88</td>
</tr>
<tr>
<td>FFS with P4P</td>
<td>0.02</td>
<td>0.55</td>
<td>0.05</td>
<td>-</td>
<td>-</td>
<td>0.02</td>
</tr>
<tr>
<td>Report Card</td>
<td>0.02</td>
<td>0.69</td>
<td>0.07</td>
<td>0.78</td>
<td>-</td>
<td>0.04</td>
</tr>
<tr>
<td>Salary</td>
<td>0.85</td>
<td>0.38</td>
<td>0.88</td>
<td>0.02</td>
<td>0.04</td>
<td>-</td>
</tr>
</tbody>
</table>

#### Table 2c: P-values from Mann Whitney U-Test on Agent's Actions for Harmful Services

<table>
<thead>
<tr>
<th>Payment Mechanisms</th>
<th>CAP</th>
<th>CAP with P4P</th>
<th>FFS</th>
<th>FFS with P4P</th>
<th>Report Card</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP</td>
<td>-</td>
<td>0.30</td>
<td>0.01</td>
<td>0.00</td>
<td>0.78</td>
<td>0.85</td>
</tr>
<tr>
<td>CAP with P4P</td>
<td>0.30</td>
<td>-</td>
<td>0.14</td>
<td>0.00</td>
<td>0.53</td>
<td>0.21</td>
</tr>
<tr>
<td>FFS</td>
<td>0.01</td>
<td>0.14</td>
<td>-</td>
<td>0.00</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>FFS with P4P</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Report Card</td>
<td>0.78</td>
<td>0.53</td>
<td>0.05</td>
<td>0.00</td>
<td>-</td>
<td>0.49</td>
</tr>
<tr>
<td>Salary</td>
<td>0.85</td>
<td>0.21</td>
<td>0.01</td>
<td>0.00</td>
<td>0.49</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 3a: P-values from Mann Whitney U-Test on Net Impact on Client by Treatment

<table>
<thead>
<tr>
<th>Payment Mechanisms</th>
<th>CAP</th>
<th>CAP with P4P</th>
<th>FFS</th>
<th>FFS with P4P</th>
<th>Report Card</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP</td>
<td>-</td>
<td>0.30</td>
<td>0.23</td>
<td>0.00</td>
<td>0.03</td>
<td>0.94</td>
</tr>
<tr>
<td>CAP with P4P</td>
<td>0.30</td>
<td>-</td>
<td>0.86</td>
<td>0.00</td>
<td>0.66</td>
<td>0.33</td>
</tr>
<tr>
<td>FFS</td>
<td>0.23</td>
<td>0.86</td>
<td>-</td>
<td>0.00</td>
<td>0.59</td>
<td>0.30</td>
</tr>
<tr>
<td>FFS with P4P</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Report Card</td>
<td>0.03</td>
<td>0.66</td>
<td>0.59</td>
<td>0.00</td>
<td>-</td>
<td>0.03</td>
</tr>
<tr>
<td>Salary</td>
<td>0.94</td>
<td>0.33</td>
<td>0.30</td>
<td>0.00</td>
<td>0.03</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 4a: P-values from Mann Whitney U-Test on Cost to Employer by Payment Type

<table>
<thead>
<tr>
<th>Payment Mechanisms</th>
<th>CAP</th>
<th>CAP with P4P</th>
<th>FFS</th>
<th>FFS with P4P</th>
<th>Report Card</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP</td>
<td>-</td>
<td>1.00</td>
<td>0.00</td>
<td>0.45</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>CAP with P4P</td>
<td>1.00</td>
<td>-</td>
<td>0.00</td>
<td>0.70</td>
<td>0.28</td>
<td>0.54</td>
</tr>
<tr>
<td>FFS</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>FFS with P4P</td>
<td>0.45</td>
<td>0.70</td>
<td>0.00</td>
<td>-</td>
<td>0.43</td>
<td>0.39</td>
</tr>
<tr>
<td>Report Card</td>
<td>0.01</td>
<td>0.28</td>
<td>0.00</td>
<td>0.43</td>
<td>-</td>
<td>0.00</td>
</tr>
<tr>
<td>Salary</td>
<td>0.04</td>
<td>0.54</td>
<td>0.00</td>
<td>0.39</td>
<td>0.00</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 4b: P-values from Mann Whitney U-Test on Benefit-to-Cost to Employer by Payment Type

<table>
<thead>
<tr>
<th>Payment Mechanisms</th>
<th>CAP</th>
<th>CAP with P4P</th>
<th>FFS</th>
<th>FFS with P4P</th>
<th>Report Card</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP</td>
<td>-</td>
<td>0.70</td>
<td>0.00</td>
<td>0.04</td>
<td>0.07</td>
<td>0.70</td>
</tr>
<tr>
<td>CAP with P4P</td>
<td>0.70</td>
<td>-</td>
<td>0.00</td>
<td>0.07</td>
<td>0.26</td>
<td>0.91</td>
</tr>
<tr>
<td>FFS</td>
<td>0.00</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>FFS with P4P</td>
<td>0.04</td>
<td>0.07</td>
<td>0.00</td>
<td>-</td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>Report Card</td>
<td>0.07</td>
<td>0.26</td>
<td>0.00</td>
<td>0.09</td>
<td>-</td>
<td>0.09</td>
</tr>
<tr>
<td>Salary</td>
<td>0.70</td>
<td>0.91</td>
<td>0.00</td>
<td>0.04</td>
<td>0.09</td>
<td>-</td>
</tr>
</tbody>
</table>
Appendix 2:

Introduction
This is an experiment in decision-making. You will have an opportunity to earn a considerable amount of cash through your participation in this experiment. You will be required to complete the experiment individually. You are not allowed to communicate with any other participant at any point during the course of the experiment. If you complete the experiment, you will be paid for the decisions you make in the task. You may withdraw from the experiment at any time. If you withdraw from the experiment before its completion, you will only be paid for the portion of the experiment that you have completed. If you do remain in the experiment, then you should feel free to try to make as much money as you can. All responses and decisions will be anonymous and the only piece of experimental material that will contain your identity will be your receipt of payment at the end of the experiment.

Before we begin, please set your cell phones to silent. We ask that you not make calls or send text messages until the experiment is complete. We also ask that you not talk to other participants in the experiment until after the experiment is complete.

Instructions

Basic Overview:
In this experiment you will be asked to proofread 10 short essays. The number of essays that you proofread and the number of corrections per essay that you complete is up to you. Your income earned in this experiment will be based on your performance in the task.

Proofreading Task:
In a previous session of this experiment we asked participants to proofread short essays for typographical and spelling errors. They were given 10 essays with 10 errors in each essay. These participants were initially given $25.00 for completing the proofreading task, but had $0.25 taken away for each error they were not able to correct. For instance:
  • If in Essay 1, the participant corrected 5 out of 10 errors, $0.25*5=$1.25 was taken away from their initial endowment.
  • If in Essay 2 the participant corrected 0 out of 10 errors, $0.25*10=$2.50 was taken away from their initial endowment.
  • A total of $3.75 was taken away from their initial endowment in this example.

Note that if the participant was unable to correct any errors, they received $0.00 for the experiment.

Your task in the experiment will be to proofread the essays that the first participants have already edited. The changes that you make to their essays will allow them to earn back a portion of their $25.00 that they lost. For each error that you correctly identify, we will
reimburse the participant in the first group $0.15. However, if you make an incorrect change the participant in the first group will lose $0.05.

The mistakes that the first participants identified will no longer be in the essay and therefore you will not need to correct them.

The mistakes that the first group missed will still be in the essay. Your task is to proofread these essays for the errors that the first group missed. You will be given assistance with your editing task. For each error in the essay, we have highlighted 3 words, only one of which is the actual error. The first group was not given this assistance in their proofreading task. For example, if the person who initially proofread the essay missed 3 errors you will see 9 highlighted words, only 3 of which contain actual errors. The total number of actual errors will be reported to you at the top of each essay.

This proofreading task allows each participant from the previous session to obtain help from one person in this part of the experiment for his or her essays.

In total you will be presented with 10 essays to edit, each from a different participant. You will have a total of 10 minutes to complete this task. At the end of the 10-minute period you will no longer be able to proofread essays.

**Proofreading Payment:**
You will be paid $0.20 for each of the highlighted words that you attempt to correct within the 10-minute time frame. You do not have to correctly change the typo to be paid, but the person whose essay you are editing will only be reimbursed for correct editing and will lose money if incorrect editing.

For instance if you make 9 changes to Essay 1 your payment will be calculated as follows:
- If you correctly identify 3 of the highlighted errors and incorrectly identify 6 of the highlighted errors:
  - You will receive $0.20 for all 9 changes and your total payment for Essay 1 will be $1.80
  - The participant whose essay you are editing will receive $0.15 for each of the 3 correctly identified errors (a total of $0.45) and lose $0.05 for the 6 incorrectly changed words (a total of $0.30), so their total reimbursement payment will be $0.45-$0.30=$0.15.

Note: Your total payment will be calculated by summing your changes over all 10 essays.

**Completion of Task:**
You will be given 10 minutes to complete the proofreading task, however if you are satisfied with the number of corrections that you have made you can leave the experiment by clicking the “Finish” button on the lower left-hand side of your screen before the 10 minutes are up. At this time the instructor will provide you with a brief survey to complete before you are paid and are free to leave.
Appendix 3:
Quiz: You will now complete a quiz to see if you understand how your earnings will be calculated.

1. If you correctly identify 5 of the highlighted errors and incorrectly identify 1 of the highlighted errors in Essay 1:
   For your corrections in Essay 1 you receive _____
   The participant whose essay you are editing will receive _____

2. If you correctly identify 1 of the highlighted errors and incorrectly identify 5 of the highlighted errors in Essay 2:
   For your corrections in Essay 2 you receive _____
   The participant whose essay you are editing will receive _____

3. If you correctly identify 7 of the highlighted errors and incorrectly identify 2 of the highlighted errors in Essay 3:
   For your corrections in Essay 3 you receive _____
   The participant whose essay you are editing will receive _____

4. If you correctly identify 4 of the highlighted errors and incorrectly identify 4 of the highlighted errors in Essay 4:
   For your corrections in Essay 4 you receive _____
   The participant whose essay you are editing will receive _____

5. If you correctly identify 4 of the highlighted errors and incorrectly identify 0 of the highlighted errors in Essay 5:
   For your corrections in Essay 5 you receive _____
   The participant whose essay you are editing will receive _____

6. If you correctly identify 0 of the highlighted errors and incorrectly identify 0 of the highlighted errors in Essay 6:
   For your corrections in Essay 6 you receive _____
   The participant whose essay you are editing will receive _____

7. If you correctly identify 9 of the highlighted errors and incorrectly identify 5 of the highlighted errors in Essay 7:
   For your corrections in Essay 7 you receive _____
   The participant whose essay you are editing will receive _____

8. If you correctly identify 1 of the highlighted errors and incorrectly identify 6 of the highlighted errors in Essay 8:
   For your corrections in Essay 8 you receive _____
   The participant whose essay you are editing will receive _____

9. If you correctly identify 3 of the highlighted errors and incorrectly identify 1 of the highlighted errors in Essay 9:
For your corrections in Essay 9 you receive _____
The participant whose essay you are editing will receive _____

10. If you correctly identify 0 of the highlighted errors and incorrectly identify 6 of the highlighted errors in Essay 10:

   For your corrections in Essay 10 you receive _____
The participant whose essay you are editing will receive _____

11. Please calculate your total earnings from correcting these ten essays (questions #1-10)
    Total Earnings _____
Appendix 4:

Instructions
This is an experiment in decision-making. You will have an opportunity to earn a considerable amount of cash through your participation in this experiment. You will be required to complete the experiment individually. You are not allowed to communicate with any other participant at any point during the course of the experiment. If you complete the experiment, you will be paid for the decisions you make in the task. You may withdraw from the experiment at any time. If you withdraw from the experiment before its completion, you will only be paid for the portion of the experiment that you have completed. If you do remain in the experiment, then you should feel free to try to make as much money as you can. All responses and decisions will be anonymous and the only piece of experimental material that will contain your identity will be your receipt of payment at the end of the experiment.

Basic Overview
In this experiment, you are asked to proofread several short essays for spelling and typographical errors. At the end of the experiment you will be paid for your performance in this proofreading task.

Earnings Task:
Each person starts the experiment with $25. You will be given 10 essays to proofread. Each essay contains 10 spelling and typographical errors. For each error you correctly identify there is no change to your earnings, however for each error in the essays that you fail to correct you will have $0.25 taken away from your $25.00. You will not be penalized for changes made to portions of the text that did not originally contain an error.

You will have a total of 50 minutes to complete the proofreading task. At the end of the 50-minute period, the instructor will collect the essays to check for correctness and calculate your earnings.
For example:
• If in Essay 1, if you correct 5 out of 10 errors, $1.25 was taken away from your $25.00.
• If in Essay 2 if you correct 0 out of 10 errors, $2.50 was taken away from your $25.00.
• A total of $3.75 would be taken away from your $25.00 in this example.

Note: total earnings will be calculated over all 10 essays.

Future Earnings:
In future sessions of this experiment, other participants will proofread your corrected essays for the spelling and typographical errors that you were unable to correct, these participants will not know whose essay they are proofreading and you will not know who proofread your essay. This will allow you to earn some of your money back. For each error that they find that you failed to correct, we will award you $0.15.
For instance:

- If you missed 10 errors and the participants in the future sessions found 2 of them, you will receive $0.30.
- For this reason it is important that we have an email address to contact you so that you collect these future earnings.