Optimal Incentives under Moral Hazard: From Theory to Practice*

George Georgiadis and Michael Powell†

October 23, 2019

Preliminary & Incomplete

Abstract

This paper aims to improve the practical applicability of the classic theory of incentive contracts under moral hazard. We show that the information provided by an A/B test of incentive contracts is a sufficient statistic for the question of how best to locally improve a status quo incentive contract, given a priori knowledge of the agent's monetary preferences. We assess the empirical relevance of this result using data from DellaVigna and Pope's (2017) study of a variety of incentive contracts. Finally, we discuss how our framework can be extended to incorporate additional considerations beyond those in the classic theory.

*We are grateful to Iwan Barankay, Dan Barron, Hector Chade, Ben Golub, Eddie Lazear, Nathan Seegert, and Jeroen Swinkels, as well as to participants at several seminars and conferences for helpful comments. Finally, we thank Henrique Brasiliense De Castro Pires for excellent research assistance.

†Kellogg School of Management, Northwestern University, Evanston, IL 60208, U.S.A., g-georgiadis@kellogg.northwestern.edu and mike-powell@kellogg.northwestern.edu .

1 Introduction

Decades of research in agency theory and organizational economics theory has been occupied with the positive question of why organizations look the way they do: Why do incentive contracts have the features they do? Why are organizations dysfunctional in the ways they are? As positive theories, they have been successful at delivering deep insights into fundamental trade-offs. But as prescriptive theories, they have been largely underwhelming. The optimal organization in a given environment often depends in complicated and subtle ways on unobservable characteristics of that environment. To bridge the gap between the positive and prescriptive requires figuring out how to make the relevant aspects of the environment observable to the relevant decision makers and to characterize optimal arrangements given the information they plausibly might be able to access.

We aim to take a small and manageable step towards a prescriptive contract theory. Instead of asking "what is the best incentive contract?", we ask a narrower question, but one that is relevant in any ongoing organization: "what is the best way to *improve upon an existing contract*?" Answering the former question requires omniscience. Answering the latter requires data. The goal of this paper is to described the kind of data that are useful for answering this question, show how to use it, and provide an empirical proof of concept.

To introduce our main ideas and to illustrate two problems that our approach has to overcome, let us consider an example. Suppose you are a manager at a company that sells kitchen knife sets. You hire teenagers each summer to sell them door to door, and you pay them a simple linear piece rate for doing so. You have access to the sales data for your workforce, and you are interested in knowing whether, and how, you should change the piece rate. Suppose your gross profit margin for selling a knife set is m, the piece rate is α, and your worker's average sales are a. Your profits are therefore $\Pi = (m - \alpha)\, a$. If you were to marginally increase your piece rate, the effect on your profits would be

$$\frac{d\Pi}{d\alpha} = (m - \alpha)\, \frac{da}{d\alpha} - a, \tag{1}$$

where the first term represents the effect on your net revenues, and the second term represents the effect on your wage bill.

You know your gross profit margin, the current piece rate, and the current average sales, but you do not know your workers' *behavioral response*, $da/d\alpha$, to an increase in the piece rate. Given observational data alone, constructing this behavioral response requires knowing a lot about the problem your workers face: How much do they like money? What are their effort costs? If they work a little harder, what is going to happen to the distribution of their

sales? These are questions you likely do not know the answer to, but importantly, they are questions you do not need to know the answer to if you are willing to run an experiment.

Suppose you decide to run an A/B test on your workforce. You randomly divide it into a treatment and a control group, you increase the piece rate by a small amount in the treatment group, and you have access to the data on the distribution of output for both the status quo contract and the test contract. You can use this data to estimate $da/d\alpha$, and you can use the above expression to determine whether you should marginally increase or decrease your piece rate.

This example teaches us two lessons. The first is that observational data is not informative enough to provide guidance for decision making in this context, just as a snapshot of price-quantity data is not informative enough for telling a manager how to change prices. The second lesson is that instead of having to know the details of the worker's unobservable characteristics, it suffices to estimate a simple behavioral response, a lesson that echoes that of the growing literature on sufficient statistics for welfare analysis.

The example also sidesteps two important issues that we will have to address. First, it restricts attention to linear contracts. This is a severe restriction, as the existing contract may not be linear, and improving upon the existing contract may well entail putting in place a nonlinear contract with features such as bonuses or accelerators with increasing piece rates. Second, it asks a local question—how best to marginally improve upon the status quo contract—and for practical applications, we are interested in non-local changes. We address each of these issues in turn.

To do so, we consider the canonical principal-agent framework under moral hazard, as in Holmström (1979). Facing a contract w, which is a mapping from output to payments received, an agent chooses an unobservable and privately costly effort level a, which determines the distribution over outputs $f(\cdot \mid a)$, which we normalize so that the mean output is a. As in Holmström (1979), we assume that the agent's first-order conditions characterize his effort choice, and we assume that his preferences over money and his effort costs are additively separable and given by $v(w) - c(a)$.

Given any status quo contract w, let us consider the effects of an arbitrary nonlinear change dw to the contract. This change directly affects the expected wage bill by $\mathbb{E}[dw]$, and leads the agent to change his effort level by some amount, da. The total effect on the principal's profits is therefore

$$d\Pi = \left(m - \int w f_a\right) da - \mathbb{E}[dw],$$

which is the appropriate generalization of (1) to nonlinear contracts. The main challenge to

figuring out the best marginal change to the status quo contract is that the agent's response da depends on dw, and there is a continuum of ways in which the contract can be changed. Our main lemma shows that, given knowledge of the agent's marginal preferences for money, the information provided by a *single* A/B test of incentive contracts (which allows the principal to estimate da for a particular dw) is a sufficient statistic for the estimation of the agent's behavioral response to *any* marginal change to the contract.

The argument for this sufficient statistic result reveals how to use the data generated by an A/B test, and so it is worth detailing informally here. Given a contract, an agent will exert effort up to the point where his marginal costs of exerting additional effort equal his *marginal incentives*, which are given by $I = cov\left(v(w), f_a/f\right)$. That is, he will work harder if doing so increases the likelihood of well-compensated outputs and decreases the likelihood of poorly compensated outputs. This condition implies that the agent's behavioral response to a change in his marginal incentives, da/dI, are independent of the change in the contract that led to the change in marginal incentives. His behavioral response to a marginal change in the contract, $\widetilde{dw}$, therefore can be expressed as $\widetilde{da} = (da/dI)\,\widetilde{dI}$. Predicting how he will respond to a change in the contract therefore requires information about how the agent will respond to a change in his marginal incentives, and it requires information about how a change in the contract affects the agent's marginal incentives. A *single* A/B test, together with knowledge of the agent's marginal preferences for money, provides all the information needed to estimate these quantities.

To make use of the information from an A/B test, consider a test contract that increases the agent's mean output. Comparing the output distributions under the status quo contract and the test contract allows us to estimate which output levels become more and less likely, identifying f_a. Given an estimate of f_a and knowledge of the agent's marginal preferences for money, we can infer how the test contract changed the agent's marginal incentives, dI, which allows us to identify the agent's behavioral response to a change in marginal incentives, da/dI. It also provides the information required to estimate how *any other* marginal change to the status quo contract affects the agent's marginal incentives, $\widetilde{dI}$, and therefore the agent's effort choice $\widetilde{da} = \frac{da}{dI}\widetilde{dI}$. A single A/B test, therefore, provides all the relevant information for predicting how the principal's expected profits will change in response to any marginal change to the status quo contract and therefore serves as a sufficient statistic for the question of how best to marginally improve upon the status quo contract. This sufficient-statistic result is our main conceptual contribution. We then show that the problem of how best to locally improve upon a status quo contract is equivalent to figuring out the direction of steepest ascent in the principal's objective, which can be determined by solving a tractable constrained maximization problem.

The second important issue that the above example sidestepped was the question of how to predict the effects of *non-local* changes to the status quo contract. We show that if the agent's effort costs are isoelastic, and f_a is independent of the agent's effort choice, then the information provided by a single A/B test provides all the information needed to predict how the principal's profits will respond to *any* change to the status quo contract. In doing so, we provide an algorithm for figuring out how to use this information to optimally revise the status quo contract.

We then explore the quantitative implications of our results using data from DellaVigna and Pope's (2017) large-scale experimental study of how a variety of different incentive schemes motivate subjects in a real-effort task. We use the data from six treatments in which subjects were motivated solely by financial incentives. In all of these treatments, subjects received a fixed wage plus a contingent payment that depended on their performance in the experiment. In four of these treatments, they received a constant piece rate for every unit of performance, and the piece rate varied across the different treatments. In the remaining two treatments, subjects received a bonus if their performance exceeded a target, and the bonus varied between these treatments. We use these data to carry out two exercises.

Our first empirical exercise asks the question of whether subjects' mean output varies in the way our model predicts with our measure of the subjects' marginal incentives. We take the data from two treatments and suppose that in one of the treatments, the subjects were on the status quo contract, and in the other, they were on the test contract. This gives us fifteen status quo-test-contract pairs. For each such pair, we predict the mean output in each of the remaining four treatments and compare it to the actual average output and compute the average absolute percentage error (APE) across these four treatments. The average APE across all fifteen contract pairs is 2.1%, and it is close to 1% for the vast majority of them.[1] Data from piece-rate contracts perform well in predicting average output in the bonus contracts and vice versa.

Our second empirical exercise explores the performance of the predicted optimal contract generated by our algorithm. We use data from all treatments to estimate the parameters of the production environment using maximum likelihood estimation. Given those parameters, we compute, as a benchmark, the optimal contract and the principal's corresponding expected profit. Then, we use data from each pair of contracts, supposing that one is the status quo contract and the other is the test contract, and we use our algorithm to construct the optimally revised contract.

Averaging across the different status quo-test-contract pairs, our optimally revised contract is estimated to achieve 97.25% of the benchmark maximum profits and 48% of the

[1] As a benchmark, the average pairwise output differences across the six treatments is around 6.5%.

potential gains over the status quo contract. In contrast, the contract that induces the agent to choose the same effort level as the status quo contract at minimum cost achieves only 95.24% of the benchmark maximum profits and less than 10% of the potential gains over the status quo contract. These results suggest that in our sample, improving upon the status quo contract is better achieved by inducing the agent to change their effort level than by just figuring out how to get them to choose the same effort at a lower cost.

Although our main results apply only to the canonical principal-agent framework of Holmström (1979), we show how our main insights extend to several enrichments of the framework. For example, we show how they extend to settings where the firm employs heterogeneous agents, to settings where the agent's effort is multidimensional, and to settings where he is motivated by factors beyond direct financial incentives; *e.g.*, the threat of firing, prestige, and so on. Finally, we establish a sufficient-statistic result for settings where the principal is constrained to choosing from a parametric class of contracts, such as linear contracts, single-bonus contracts, or option contracts.

This paper straddles the theoretical and the empirical literatures on principal-agent problems under moral hazard. The canonical model (*e.g.*, Mirrlees (1976) and Holmström (1979)) considers a principal who wants to motivate an agent to choose a particular unobservable action hard. To do so, she offers a contract, which specifies a schedule of payments conditional on the realization of a signal that is correlated with the agent's action. Extensions of this model include settings in which the signal is not contractible, the agent's action is multidimensional and some tasks are easier to measure than others, or the principal and the agent interact repeatedly—see Bolton and Dewatripont (2005) for a comprehensive treatment. The goal of the theoretical literature, typically, is to characterize an optimal contract under the premise that the principal has perfect knowledge of all relevant parameters of the model.[2]

The empirical literature can be classified into (at least) two groups. The first examines the degree to which workers respond to incentives as predicted by the theory. For example, Lazear (2000) finds that the switch from hourly wages to piece-rate pay at Safelite Auto Glass led to a 44% increase in productivity, approximately half of which is attributable to workers exerting more effort, while the other half is due to selection, that is, more productive workers joining the firm and less productive ones leaving. In similar vein, Shearer (2004) finds a 20% increase in productivity when tree planters in British Columbia were paid according to piece rates, compared to hourly wages. See also Paarsch and Shearer (1999) for a related study.[3] Others study work on more complex tasks that are amenable to the multitasking

[2]One exception is Chade and Swinkels (2019), who studies a principal-agent problem under both moral hazard and adverse selection, where the principal knows all but one payoff-relevant parameters of the model.

[3]Oettinger (2001) and Fehr and Goette (2007) finds a positive effect of commissions on sales for stadium vendors and on productivity for bicycle messengers in Zürich, respectively. Bandiera et al. (2007) and

problem; see, for example, Holmström and Milgrom (1991). For example, Gibbs et al. (2017) exploits a field experiment at an Indian technology firm to estimate the impact of financial incentives for submitting ideas for process improvements. They find that incentives led employees to submit fewer but higher-quality ideas.[4] On a broader scale, Prendergast (2014) uses estimates for the elasticity of income to marginal tax rates (see, for example, Brewer, Saez and Shephard (2010)) to establish an upper bound for the responsiveness of worker productivity to incentives. The second category investigates the extent to which observed contracts are consistent with theoretical models. See, for example, Prendergast (1999) and Chiappori and Salanié (2003).

A limitation of the theoretical literature is that it often assumes omniscience on the principal's behalf (*i.e.*, she is assumed to know the agent's preferences, the actions at his disposal and the associated cost, and how these actions map into the contractible signal). On the other hand, the empirical literature usually focuses on estimating how different incentive vehicles affect performance. The goal of this paper is to bridge these literatures by exploring how an organization can combine lessons from the theoretical agency literature together with estimates such as those described above to improve its incentive system.

Our work is conceptually related to papers that use a variational approach to characterize optimal mechanisms in terms of the relevant elasticities. For example, the Lerner index relates the optimal monopoly price to the price elasticity of demand (see, for example, Tirole (1988)), and Wilson (1993) characterizes an optimal quantity-discount price-menu. Saez (2001) and a growing literature derives optimal income tax formulas using elasticities of earnings with respect to tax rates.

2 Model

Environment.— We consider the contractual relationship between a principal and one or more homogeneous agents. The principal offers an output-contingent contract $w(x)$ to each agent, who, after observing the contract, chooses effort $a \geq 0$, which is not contractible. His output, $x \in \mathbb{R}$, is realized according to some cumulative distribution function, $F(x|a)$, with probability density function (hereafter pdf) $f(x|a)$, which we assume is twice differentiable in a. Finally, payoffs are realized and the game ends. Without loss of generality, we normalize

Bandiera et al. (2009) measure the effect of introducing performance pay for managers on their subordinates' productivity. Guiteras and Jack (2018) studies the incentive effect on productivity and selection for labor workers in rural Malawi. Hill (2019) estimates the effect of an increase in the minimum wage on productivity for strawberry pickers in California.

[4]Similarly, Balbuzanov et al. (2017) finds that the introduction of incentives led journalists in Kenya to submit fewer, higher quality articles. Hong et al. (2018) estimates the impact of piece rates at a Chinese manufacturing firm on the quantity and quality of output.

a such that $a = \mathbb{E}[x|a]$, so that the agent's effort can be interpreted as his expected output.

Actions.— The principal chooses a contract $w : \mathbb{R} \to \mathbb{R}$, which is an upper-semicontinuous (hereafter, u.s.c) mapping from output to transfers made to the agent. We assume that, to ensure participation, the principal restricts attention to contracts that leave each agent with at least as much expected utility as some (generic) *status quo* contract, w_A.[5] After observing the contract, each agent chooses an effort $a \geq 0$.

Information.— Each agent knows all parameters that are pertinent to his decisions, that is, he knows his utility function $v(\omega)$, his cost function $c(a)$, and the pdf $f(\cdot|a)$ for every feasible effort level. The principal knows her marginal profit $m > 0$, and the distribution of output corresponding to two contracts, w_A and w_B. Put differently, letting $a(w)$ denote the effort induced by contract w, the principal knows the pdf's $f(\cdot|a(w_A))$ and $f(\cdot|a(w_B))$. Additionally, we assume that the principal knows $f_a(\cdot|a(w_A)) := df(\cdot|a(w_A))/da$. In practice, assuming that $a(w_A)$ and $a(w_B)$ are sufficiently close but distinct from each other, she would approximate

$$f_a(x|a(w_A)) \simeq \frac{f(x|a(w_B)) - f(x|a(w_A))}{a(w_B) - a(w_A)}. \tag{2}$$

For now, we abstain from specifying the principal's knowledge about other parameters. When convenient, we shall suppress the argument x in functions, and abbreviate $\widehat{a} = a(w_A)$, $\widehat{f} \equiv f(\cdot|a(w_A))$ and $\widehat{f_a} \equiv f_a(\cdot|a(w_A))$.

Preferences.— If an agent is paid ω and exerts effort a, then he obtains utility $v(\omega) - c(a)$, where $v : \mathbb{R} \to \mathbb{R}$ and $c : \mathbb{R}_+ \to \mathbb{R}_+$ are twice continuously differentiable, and satisfy $v'' < 0 < v'$ and $c', c'' > 0$. The agent chooses his effort to maximize his expected utility. If an agent generates output x and is paid $w(x)$, then the principal's profit is $mx - w(x)$.

The principal's objective, which we formalize in Section 3.2, is loosely speaking, to find a contract that increases her profit (relative to w_A and w_B) by as much as possible given the information at her disposal. The spirit of the exercise we consider is that the principal has data corresponding to two different contracts (*e.g.,* an A/B test), and is searching for a new contract that increases her profit by as much as possible. By analyzing this problem, one goal is to determine what (additional) information the principal must have in order to make that determination.

[5] When firms revise their performance pay plans, workers are often suspicious about the principal's intentions, which can lead to opposition (*e.g.*, in the form of unionization) and attrition; see, for example, Hall et al. (2000). Restricting attention to contracts that make workers at least as well off as a status quo contract may ease those tensions.

2.1 Benchmark

In this section, we present a benchmark, due to Holmström (1979). The canonical principal-agent model under moral hazard is formulated as a two-stage optimization problem (Grossman and Hart, 1983). In the first-stage, the principal solves for every feasible effort level, the following constrained maximization program:

$$\begin{aligned} \Pi(a) := \max_{w(\cdot)} \quad & \int [mx - w(x)]\, f(x|a)dx \qquad (\mathrm{P}_H) \\ \text{s.t.} \quad & \int v(w(x))f(x|a)dx - c(a) \geq \underline{u} \\ & a \in \arg\max_{\widetilde{a}\geq 0} \left\{ \int v(w(x))f(x|\widetilde{a})dx - c(\widetilde{a}) \right\} \end{aligned}$$

where $\underline{u}$ is the agent's outside option, which in our setting, corresponds to the agent's expected payoff from the status quo contract, w_A. The first constraint mandates that the contract gives the agent no less than $\underline{u}$ utils in expectation, while the second ensures that effort a is incentive compatible. To solve this program, one typically replaces the incentive constraint with the corresponding first-order condition, $\int v(w(x))f_a(x|a)dx = c'(a)$—see Jewitt (1988) for conditions such that doing so is without loss of generality, and the principal's choice variable, $w(\cdot)$, with $V(\cdot) \equiv v(w(\cdot))$, thus transforming (P_H) into a convex program. In the second stage, the principal solves $\Pi^* = \max_a \{\Pi(a)\}$ to find the profit-maximizing effort and the corresponding optimal contract. The second-stage problem is notoriously ill-behaved, and concave in a only under stringent conditions (Jewitt, Kadan, and Swinkels, 2008), so it is typically solved using line search.

To solve this problem, the principal must know (or make assumptions for) all payoff-relevant parameters; *i.e.,* the agent's utility function $v(\cdot)$ and his cost function $c(\cdot)$, his outside option $\underline{u}$, as well as the pdf $f(\cdot|a)$ and its derivative $f_a(\cdot|a)$ for every feasible level of effort. In many settings, it is unrealistic to expect that the principal has this information at her disposal. Motivated by this observation, we pursue a more modest objective: Given knowledge about the distribution of output corresponding to two contracts, how can the principal best improve upon them, and what additional information is necessary to do so.

3 Optimal Perturbations

In this section, we propose a methodology for finding an optimal perturbation of the status quo contract, w_A. To do so, the principal must predict the agent's response to *any* change in the offered incentives. Our key observation is that if (but only if) the principal knows or

equivalently, takes a stance on the agent's marginal utility function, $v'(\cdot)$, then she can use the data contained in her A/B test to make the needed inference. Knowledge of v', together with an envelope condition also enables the principal to restrict attention to perturbations that make the agent at least as well of as w_A.

3.1 Agent's Problem

We assume that the first-order approach is valid, so that given some contract w, the agent's optimal choice of effort, $a(w)$ satisfies the first-order condition

$$\int v(w(x))f_a(x|a(w))dx = c'(a(w))\,. \tag{IC}$$

For any u.s.c function $t : \mathbb{R} \to \mathbb{R}$, consider the family of contracts $\{w_A + \theta t\}_{\theta \geq 0}$. We shall call t a perturbation, and $w_A + \theta t$ a perturbation of the status quo contract, w_A. Define the Gateaux derivative $\mathcal{D}a(w_A, t) := da(w_A + \theta t)/d\theta|_{\theta=0}$, which exists, because w_A and t are u.s.c, and $f(\cdot|a)$ and $c(\cdot)$ are twice-differentiable with respect to a by assumption. This derivative should be interpreted as the marginal change in the agent's effort when w_A is perturbed in the direction of $w_A + t$.[6] Using (IC), it can be written in terms of primitives as

$$\mathcal{D}a(w_A, t) = \frac{\int tv'(w_A)\widehat{f}_a dx}{c''(\widehat{a}) - \int v(w_A(x))f_{aa}(x|\widehat{a})dx}\,. \tag{3}$$

Throughout the remainder of this section, we make the following assumption.

Assumption 1. *The principal knows* $\mathcal{D}a(w_A, w_B - w_A)$.

In practice, the principal would approximate $\mathcal{D}a(w_A, w_B - w_A) \simeq a(w_B) - \widehat{a}$, which is a valid approximation as long as $\|w_B - w_A\|$ is sufficiently close to zero.

3.2 Principal's Problem

The principal's expected profit from offering contract w,

$$\pi(w) := ma(w) - \int w(x)f(x|a(w))dx\,, \tag{4}$$

where $a(w)$ solves (IC).

[6] Notice that $w_A + \theta t = (1-\theta)w_A + \theta(w_A + t)$ and because the derivative is evaluated at $\theta = 0$, it represents the marginal change of effort in a neighborhood around w_A.

Suppose that w_A is replaced by $w_A + \theta t$, for some u.s.c $t : \mathbb{R} \to \mathbb{R}$. For θ sufficiently close to zero, we have

$$\pi(w_A + \theta t) \simeq \pi(w_A) + \theta \mathcal{D}\pi(w_A, t)\,, \tag{5}$$

where the Gateaux derivative $\mathcal{D}\pi(w_A, t) := d\pi(w_A + \theta t)/d\theta|_{\theta=0}$ represents the principal's marginal benefit from perturbing the contract w_A in the direction of $w_A + t$. It exists for the same reasons as $\mathcal{D}a(w_A, t)$, and using (4), it can be rewritten as

$$\mathcal{D}\pi(w_A, t) = \left(m - \int w_A \widehat{f}_a dx\right) \mathcal{D}a(w_A, t) - \int t\widehat{f} dx\,. \tag{6}$$

This expression has an analogous interpretation as (1): Perturbing the status quo contract has two effects on the principal's profit. First, it induces a change in the agent's effort, as captured by the first term, and holding effort fixed, it affects profits mechanically, as captured by the second term.

Observe that for fixed θ, maximizing (5) is equivalent to maximizing (6). Thus, we take the principal's objective to be to choose a perturbation, t, that maximizes (6) subject to the constraint that the perturbed contract, $w_A + \theta t$, gives the agent at least as much expected utility as w_A. The set of feasible perturbations must satisfy

$$\frac{d}{d\theta} \int v\left(w_A + \theta t\right) f(x|a(w_A + \theta t))dx - c(a(w_A + \theta t))\Big|_{\theta=0} \geq 0 \Leftrightarrow \int t v'(w_A)\widehat{f} dx \geq 0\,. \tag{7}$$

That is, for any feasible perturbation t, each agent's expected utility must be non-decreasing as w_A is perturbed in the direction of $w_A + t$, and we used (IC) to obtain the second expression.

To find solve this problem, the principal must be able to evaluate $\mathcal{D}a(w_A, t)$ for every (feasible) perturbation t. Observe that t appears only in the numerator of (3), so if the principal knows the agent's marginal utility function, $v'(\cdot)$, then she can use her (assumed) knowledge $\mathcal{D}a(w_A, w_B - w_A)$ to compute the denominator of (3), which in turn, will allow her to compute $\mathcal{D}a(w_A, t)$ for any other t. Moreover, knowledge of v' also allows the principal to inspect whether t satisfies (7), and hence solve the problem at hand. The following remark summarizes.

Remark 1. *For any u.s.c $t : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathcal{D}a(w_A, t) = \frac{\mathcal{D}a(w_A, w_B - w_A)}{\int \left(w_B - w_A\right) v'(w_A)\widehat{f}_a dx} \int t v'(w_A)\widehat{f}_a dx\,. \tag{8}$$

If principal knows the agent's marginal utility function, v', and Assumption 1 holds, then she

can evaluate (6) and (7) for every t.

Faced with w_A, the agent chooses his effort by equating its marginal benefit, $M(w_A) := \int v(w_A(x))\widehat{f}_a dx$, to its marginal cost. When w_A is perturbed in the direction of $w_A + t$, this marginal benefit changes at rate $\mathcal{D}M(w_A, t) = \int tv'(w_A)\widehat{f}_a dx$, which in turn, induces the agent to change his effort. Locally, this relationship is linear; that is, $\mathcal{D}a(w_A, t) = C \times \mathcal{D}M(w_A, t)$ for some constant C. Given v' and $\widehat{f}_a$, the principal can predict $\mathcal{D}M(w_A, t)$ for any t, pin down $C = \mathcal{D}a(w_A, w_B - w_A)/\mathcal{D}M(w_A, w_B - w_A)$, and compute $\mathcal{D}a(w_A, t)$ for any t.

Insofar, we have shown that to evaluate (6) and (7), it suffices that the principal knows v'. Is it also necessary? Strictly speaking, no: If, for example, the principal knows $c'(\widehat{a})$ and assumes that v' belongs to a one-parameter family of functions, then she can use (IC) to solve for the unknown parameter. Alternatively, if she knows $\mathcal{D}a(w_A, w_C - w_A)$ for some contract w_C, then she can use (8) to solve for the unknown parameter in v'. Notice, however, that in both cases, the object of interest is v'.

Observe that both (6) and (7) are linear in t. Thus, the principal can increase her objective without bound by making $t(x)$ arbitrarily large for some x, and arbitrarily small for all other x. When faced with this issue in optimization problems, a common approach is to normalize the *length* of t by imposing the constraint $\|t\| \leq 1$, where $\|\cdot\|$ is the Euclidean norm (see, for example, Section 9.4 in Boyd and Vandenberghe (2004)).[7] Imposing this constraint, and using (8), the principal's problem can be expressed as

$$\begin{aligned} \max_{t \text{ u.s.c}} \ & \mu \int tv'(w_A)\widehat{f}_a dx - \int t\widehat{f} dx \\ \text{s.t.} \ & \int tv'(w_A)\widehat{f} dx \geq 0 \\ & \int t^2 dx \leq 1 \end{aligned} \tag{P_{local}}$$

where

$$\mu = \frac{\left(m - \int w_A \widehat{f}_a dx\right) \mathcal{D}a(w_A, w_B - w_A)}{\int (w_B - w_A) v'(w_A) \widehat{f}_a dx}. \tag{9}$$

[8] This is a convex optimization program, and it can be solved using standard techniques.

[7] Alternatively, one can take $\|\cdot\|$ to be any L^p norm with $p \geq 2$. We focus on the case $p = 2$ for expositional convenience.

[8] Notice that the denominator of (3) is the negative of the second derivative of the agent's expected utility with respect to a. Therefore, one can inspect whether the first-order approach is locally valid at $\widehat{a}$, which is necessary for the validity of this analysis, by verifying that $\mathcal{D}a(w_A, w_B - w_A)$ and $\int (w_B - w_A) v'(w_A) \widehat{f}_a dx$ have the same sign. In that case, μ has the same sign as $m - \int w_A \widehat{f}_a dx$.

The following proposition characterizes the uniquely optimal perturbation, t^*.

Proposition 1. *The status quo contract, w_A, is locally optimal if and only if*

$$\lambda + \mu \frac{\widehat{f}_a}{\widehat{f}} = \frac{1}{v'(w_A)} \tag{10}$$

for all x, where $\lambda = \int \widehat{f}/v'(w_A)dx$ and μ is given in (9).

Otherwise, the optimal perturbation

$$t^* = K \times \left[\lambda v'(w_A)\widehat{f} + \mu v'(w_A)\widehat{f}_a - \widehat{f}\right], \tag{11}$$

where μ is given in (9), and $K > 0$ and $\lambda \geq 0$ are given in the proof of Proposition 1, in Appendix B.

The first part of this result, equation (10), is familiar from the canonical principal-agent model under moral hazard (Holmström, 1979), also presented in Section 2.1, and serves here the role of a consistency check.[9] Turning to (11), marginally increasing payments at x has three effects on the principal's profit: First, it relaxes the constraint (7), which has implicit value $\lambda v'(w_A)\widehat{f}$. Second, it affects the agent's effort, which has implicit value $\mu v'(w_A)\widehat{f}_a$, and third, holding effort constant, it reduces the principal's profit at rate $\widehat{f}$. Thus, the optimal perturbation increases payments at every output in proportion to the principal's net marginal benefit of doing so.

Proposition 1 is useful in that it sheds light on the informational requirements for finding an optimal perturbation. In particular, to solve (P_{local}), the principal must know (or estimate, or take a stance on) the following parameters: *(i)* the distribution of output corresponding to some effort, $\widehat{f} \equiv f(\cdot|a(w_A))$, *(ii)* the rate at which this distribution changes due to a marginal change in effort, $\widehat{f}_a \equiv f_a(\cdot|a(w_A))$, *(iii)* the Gateaux derivative $\mathcal{D}a(w_A, w_B - w_A)$ for two contracts such that $a(w_A) \neq a(w_B)$, and *(iv)* the agent's marginal utility function, $v'(\cdot)$. Moreover, it can be used to infer what assumptions can rationalize the status quo contract being optimal. For instance, consider a principal who does not know $\mathcal{D}a(w_A, w_B - w_A)$ or $\widehat{f}_a$.[10]

[9] Note however, that the dual multipliers λ and μ in Proposition 1 are given in closed form and they contain *information* not contained in the standard solutions. In particular, it is well-known that for any fixed effort level, the profit-maximizing contract satisfies (10) for *some* dual multipliers λ and μ. These multipliers are chosen such that there exists no perturbation that increases the principal's profit, while holding the agent's utility and his optimal effort choice constant (Jewitt, Kadan, and Swinkels, 2008). In contrast, the multipliers characterized in Proposition 1 also consider perturbations in which the agent changes his effort level.

[10] In practice, to estimate these quantities, a firm must experiment by offering a contract other than w_A to its workers. Lincoln Electric, for example, is infamous for its use of piece rates with factory workers, and the fact that it does not experiment with its performance pay plans (in fear of ratchet effects) (Hall et al., 2000).

Nevertheless, she can use (10) to reverse-engineer what effort responses or assumptions about the agent's marginal utility function are consistent with w_A being optimal, and evaluate the extent to which they are reasonable.

While Proposition 1 can be used to draw qualitative insights about profitable perturbations, its value in quantifying the optimal perturbation is limited. Given t^*, the principal should replace the status quo contract with $w(\cdot) \equiv w_A(\cdot) + \theta t^*(\cdot)$ for some step size $\theta > 0$ *close to zero*. This is important, as t^* is an optimal perturbation only for θ sufficiently small—in computing (3), (5), and (7), we ignored terms of order θ^2, and doing so is valid only if $\theta \simeq 0$. However, as the projected increase in profits is of order θ, and because practically changing incentives involves discrete costs, it likely makes sense for the principal to replace w_A only if θ is bounded away from zero. Using the methodology developed in this section (and in particular, leveraging Remark 1), we turn to this objective in the next section.

4 An Approximate Algorithm

In this section, we develop an algorithm for finding an optimal perturbation of w_A without the restriction that θ is *small*. Then in Section 5 we test its performance using a dataset from DellaVigna and Pope (2017). Towards this goal, we make the following two assumptions:

Assumption 2. *For all a in some interval that contains $\widehat{a}$, $f_a(\cdot|a) \equiv \widehat{f_a}(\cdot)$.*

This assumption allows the principal to predict the distribution of output corresponding to effort levels other than $\widehat{a}$, and it implies that the marginal incentive of effort corresponding to contract w,

$$M(w) = \int v(w(x)) f_a(x|a(w)) dx = \int v(w) \widehat{f_a}\,,$$

does not depend on a itself.

Assumption 3. *The agent's cost function is of the form $c'(a) = c_0 a^{1/\epsilon}$ for some constants $c_0 > 0$ and $\epsilon \geq 0$; i.e., the agent faces isoelastic costs of effort.*

The implication of this assumption is that for any contract w, effort $a(w)$, and marginal incentives, $M(w)$, are related by

$$\log a(w) = \beta + \epsilon \log M(w)\,, \tag{12}$$

where β and ϵ are parameters to be determined. Using $a(w_i)$ and the estimated $M(w_i)$

(together with Assumption 2) for $i \in \{A, B\}$, we have

$$\epsilon = \frac{\log\left(a(w_A)/a(w_B)\right)}{\log\left(\int v(w_A)\widehat{f}_a / \int v(w_B)\widehat{f}_a\right)} \quad \text{and} \quad \beta = \log a(w_A) - \epsilon \log \int v(w_A)\widehat{f}_a \,. \tag{13}$$

Notice that $c_0 = e^{-\beta/\epsilon}$. This assumption enables the principal to predict the agent's effort corresponding to any contract w.[11]

Suppose that Assumptions 2 and 3 hold. Then the principal's problem is expressed by the following constrained maximization program:

$$\max_{w(\cdot),\,\Delta a} \; m(\widehat{a} + \Delta a) - \int w(\widehat{f} + \Delta a \widehat{f}_a) dx \tag{$\widehat{P}$}$$

$$\text{s.t.} \int v(w)\widehat{f}_a dx = \left(\frac{\widehat{a} + \Delta a}{\widehat{a}}\right)^{1/\epsilon} \int v(w_A)\widehat{f}_a dx \tag{$\widehat{IC}$}$$

$$\int v(w)(\widehat{f} + \Delta a \widehat{f}_a) dx \geq \int v(w_A)\widehat{f} dx + \frac{\epsilon e^{-\beta/\epsilon}}{1+\epsilon}\left[(\widehat{a} + \Delta a)^{\frac{1+\epsilon}{\epsilon}} - \widehat{a}^{\frac{1+\epsilon}{\epsilon}}\right] \tag{$\widehat{IR}$}$$

where ϵ is given in (13), and Δa represents the change in effort relative to $\widehat{a}$ that the principal aims to induce with the contract w. Let us explain $(\widehat{P})$, $(\widehat{IC})$, and $(\widehat{IR})$.

Suppose that the principal wants to choose a contract w that motivates the agent to choose effort $a(w) = \widehat{a} + \Delta a$. Then using (12) and (13), and re-arranging terms, it follows that w must satisfy $(\widehat{IC})$.

Recall that by assumption, the principal restricts attention to contracts that give the agent no less expected utility than the status quo contract. This constraint can be written as

$$\int v(w(x)) f(x|\widehat{a} + \Delta a) dx - c(\widehat{a} + \Delta a) \geq \int v(w_A)\widehat{f} - c(\widehat{a}) \,,$$

or equivalently, as $(\widehat{IR})$, using that $f(\cdot|\widehat{a} + \Delta a) = \widehat{f} + \Delta a \widehat{f}_a$ by Assumption 2, and $c(\widehat{a} + \Delta a) - c(\widehat{a})$ is equal to the right-hand side of $(\widehat{IR})$ by Assumption 3.

Finally, the principal's profit, $\pi(w) = m(\widehat{a} + \Delta a) - \int w(x) f(x|\widehat{a} + \Delta a) dx$ can be rewritten as $(\widehat{P})$ using Assumption 2.

This program should be interpreted as an approximation to the optimal contracting problem given in Section 2.1. It can be solved using the standard two-step approach proposed by Grossman and Hart (1983): In the first stage, one fixes a Δa and solves $(\widehat{P})$ subject to $(\widehat{IC})$

[11] Alternatively, (12) can be replaced by the linear relationship, $a(w) = \beta_0 + \beta_1 M(w)$, where β_0 and β_1 are estimated using $\{a(w_i), M(w_i)\}$ for $i \in \{A, B\}$. This model is equivalent to assuming that the agent's cost function has constant unit elasticity. If $\|w - w_A\| \simeq 0$, then it is also equivalent to (12). Otherwise however, the two models typically generate drastically different predictions for the agent's effort. We evaluate both models in Section 5.1, and find that (12) outperforms the linear model.

and $(\widehat{IR})$ to find the profit-maximizing contract that is projected to lead to effort $\widehat{a}+\Delta a$. Let us denote the objective of this program by $\widehat{\Pi}(\widehat{a}+\Delta a)$. In the second stage, one solves

$$\widehat{\Pi}^* = \max_{\Delta a} \widehat{\Pi}(\widehat{a}+\Delta a)\,. \tag{14}$$

We make three remarks: First, the informational requirements for solving this problem are the same as those for finding an optimal perturbation, described in Remark 1: the principal must know the pdf corresponding to $a(w_A)$, $\widehat{f} \equiv f(\cdot|a(w_A))$ and its derivative $\widehat{f_a} \equiv f_a(\cdot|a(w_A))$, and the agent's marginal utility function, v'.[12] Second, if $\Delta a = 0$, the solution to the first-stage problem is equivalent to solving (P_H); *i.e.,* $\widehat{\Pi}(\widehat{a}) = \Pi(\widehat{a})$. If $\Delta a = a(w_B) - \widehat{a}$, then given Assumptions 2-3, we have $\widehat{\Pi}(a(w_B)) = \Pi(a(w_B))$.

Finally, notice that the first-stage problem can be transformed into a convex program by changing the principal's choice variable, $w(\cdot)$, to $V(\cdot) \equiv v(w(\cdot))$ only if $\widehat{f} + \Delta a \widehat{f_a} \geq 0$ for all x. This, therefore, imposes a constraint on the set of Δa's that the principal can consider. Alternatively, one can linearize the constraints by using the approximation $v(w(x)) \simeq v(w_A(x)) + (w(x) - w_A(x))v'(w_A(x))$. Then it is convenient to use the transformation $t \equiv w - w_A$, so that $(\widehat{P})$ can be expressed as

$$\begin{aligned}
&\max_{t(\cdot),\Delta a}\ m(\widehat{a}+\Delta a) - \int (w_A + t)(\widehat{f} + \Delta a \widehat{f_a})dx && (\widehat{P}_{lin})\\
&\text{s.t. } \int t v'(w_A)\widehat{f_a}dx = \left[\left(\frac{\widehat{a}+\Delta a}{\widehat{a}}\right)^{1/\epsilon} - 1\right]\int v(w_A)\widehat{f_a}dx\\
&\qquad \int t v'(w_A)\left(\widehat{f} + \Delta a \widehat{f_a}\right)dx \geq \frac{\epsilon e^{-\beta/\epsilon}}{1+\epsilon}\left[(\widehat{a}+\Delta a)^{\frac{1+\epsilon}{\epsilon}} - \widehat{a}^{\frac{1+\epsilon}{\epsilon}} - \frac{1+\epsilon}{\epsilon}\widehat{a}^{1/\epsilon}\Delta a\right],
\end{aligned}$$

where we used $\int v(w_A)\widehat{f_a} = c'(\widehat{a}) = e^{-\beta/\epsilon}\widehat{a}^{1/\epsilon}$ to obtain the expression for the second constraint. In this case, observe that the first-stage problem is linear in t, and so the objective can be increased without bound by making $t(x)$ arbitrarily large for some x's, and arbitrarily small otherwise. To ensure that a solution exists, one typically normalizes the *length* of t by imposing the constraint $\|t\| \leq C$ for some constraint C. We denote the objective of the first-stage problem corresponding to $(\widehat{P}_{lin})$ given some Δa by $\widehat{\Pi}_{lin}(\widehat{a}+\Delta a)$, and the solution of the corresponding second-stage problem, $\widehat{\Pi}^*_{lin} = \max_{\Delta a} \widehat{\Pi}_{lin}(\widehat{a}+\Delta a)$.

[12] Notice that v (instead of v') appears in $(\widehat{IC})$ and $(\widehat{IR})$. Nevertheless, because $\int \widehat{f_a} = 0$ and $\int v\widehat{f}$ appears on both sides of $(\widehat{IR})$, it follows that it suffices that the principal knows (or takes a stance on) v'.

Contract (in ¢)	Mean # of keystrokes	Std. Errors	N
$w_1(x) = 100$	1521	31.23	540
$w_2(x) = 100 + 0.001x$	1883	28.61	538
$w_3(x) = 100 + 0.01x$	2029	27.47	558
$w_4(x) = 100 + 0.04x$	2132	26.42	566
$w_5(x) = 100 + 0.10x$	2175	24.28	538
$w_6(x) = 100 + 40\,\mathbb{I}_{\{x \geq 2000\}}$	2136	24.66	545
$w_7(x) = 100 + 80\,\mathbb{I}_{\{x \geq 2000\}}$	2188	22.99	532

Table 1: Monetary incentive treatments in the experiment of DellaVigna and Pope (2017)

5 Empirical Validation

The goal of this section is to test the predictions of our model and to illustrate how one can apply the techniques developed in the previous section. To do so, we use a dataset from DellaVigna and Pope (2017), who present the findings from a real-effort experiment conducted on Amazon MTurk, in which subjects were tasked with alternating 'a' and 'b' keystrokes during a ten-minute interval. Table 2 summarizes seven of the incentive treatments that the authors implemented and are relevant for our purposes, where x denotes the number of 'a-b' keystrokes (during the 10-minute interval) and N denotes the sample size corresponding to each treatment. As an example, the third contract, $w_3(x) = 100 + 0.01x$, specifies that a subject receives \$1 irrespective of his performance, plus 0.01¢ for each 'a-b' keystroke. Each subject was randomly assigned to a single treatment, and undertook the button-pressing task once. During the course of a treatment, subjects could see the treatment characteristics (*i.e.,* the incentives offered), a count-down clock, as well as the number of keystrokes and the accumulated earnings at every moment on their computer screen.[13]

We use this dataset to perform the following two exercises. A necessary condition for finding the optimal perturbation is that the model can accurately predict how a change in the contract affects effort, and consequently the principal's profit. Exploiting the fact that this dataset contains seven different treatments, in Section 5.1, we pick any pair of treatments, consider them to be the principal's A/B test, and use them to predict effort and the principal's profit for each of the other treatments (assuming a fictitious value for the principal's marginal profit per 'a-b' keystroke, m). We also evaluate the accuracy of each prediction as a function of our assumptions about the subjects' (common) marginal utility function.

[13]To be specific, each subject was awarded a *point* for every 100 'a-b' keystrokes, and the payment was a function of the points accumulated. For example, under the third contract, a subject would receive \$1 irrespective of his or her performance, plus a cent for every point accumulated. To simplify the exposition, we take the performance measure to be the number of 'a-b' keystrokes instead of the number of points accumulated over the 10-minute interval.

In Section 5.2, again taking any pair of treatments to constitute the principal's A/B test, we solve $(\widehat{P})$ and $(\widehat{P}_{lin})$ to find an optimal perturbation of w_A. Then, using all seven treatments, we estimate the parameters of the model, which we use, first, to find the optimal contract by solving (P_H), second, to counterfactually estimate the profitability of the optimal perturbations corresponding to each A/B test, and third, to compare it to the optimal contract.

We now discuss two aspects of this setting that differ from our model, and can adversely affect its performance. The first is that in our model, the agent chooses effort once and for all, whereas in the experiment, subjects can adjust the intensity of their effort over time. Second, while it is likely that subjects differ in their ability to perform in this task, or in their willingness to do so, or in other dimensions, because each subject participated once in a single treatment, we cannot classify them into different types absent additional assumptions. As such, we treat subjects as being homogeneous, and use our baseline model instead of the one presented in Appendix A.1. To be specific, we assume that at the outset of the experiment, each subject observes the offered contract, and chooses "effort" a. Then the number of 'a-b' keystrokes that he or she accomplishes during the 10-minute interval, x, is drawn from some probability distribution with expected value a. Thus, the effort chosen by each subject faced with a given treatment is equal to the respective mean number of 'a-b' keystrokes.

Let us begin by selecting an arbitrary pair of the treatments listed in Table 1, label them w_A and w_B, and suppose that the principal has output data for these two treatments only. (There are 21 such pairs.) Setting $a(w_A)$ and $a(w_B)$ equal to the respective mean number of 'a-b' keystrokes, we can estimate (*e.g.,* using the kernels) the corresponding pdf's of output, $\widehat{f} = f(\cdot|a(w_A))$ and $f(\cdot|a(w_B))$, and approximate

$$\widehat{f}_a(x) \simeq \frac{f(x|a(w_B)) - f(x|a(w_A))}{a(w_B) - a(w_A)}$$

for every x. Next, we assume a particular marginal utility function v'. Let us assume that each subject's utility function exhibits constant relative risk aversion (CRRA) with (common) coefficient $\rho \in [0, 1)$, and so $v'(\omega) = \omega^{-\rho}$. Finally, we assume a particular value for m, the principal's marginal profit per 'a-b' keystroke.

5.1 Predicting Effort and Profits

The goal of this section is to evaluate the ability of the models presented in Section 4 to predict each subject's effort and the principal's profit. For any available A/B test, we will use our model to predict effort and profits for each of the other treatments, and then compare

our predictions to the actual values.

Using Assumption 2, for any contract w, the corresponding marginal benefit of effort $M(w) = \int w^{1-\rho} \widehat{f}_a dx \,/\, (1-\rho)$.[14] Then it follows from Assumption 3 and (12) that

$$a(w) = \left[\frac{\int w^{1-\rho} \widehat{f}_a dx}{\int w^{1-\rho} \widehat{f}_a dx} \right]^{\epsilon} a(w_A) \text{ , where } \epsilon = \frac{\log\left(a(w_A)/a(w_B)\right)}{\log\left(\int w_A^{1-\rho} \widehat{f}_a dx / \int w_B^{1-\rho} \widehat{f}_a dx\right)} \tag{15}$$

and $\beta = \log\left(a(w_A)\right) - \epsilon \log\left(\int w_A^{1-\rho} \widehat{f}_a dx/(1-\rho)\right)$.[15] In addition, to the "logarithmic model" described above, we also present the predictions from the "linear model" described in footnote 15. In this case, for any contract w, the model predicts

$$a(w) = a(w_A) + [a(w_B) - a(w_A)] \, \frac{\int (w^{1-\rho} - w_A^{1-\rho}) \widehat{f}_a dx}{\int (w_B^{1-\rho} - w_A^{1-\rho}) \widehat{f}_a dx} \,. \tag{16}$$

We denote the effort prediction corresponding to w_i using (15) and (16) by $\widehat{a}_{lin}(w_i)$ and $\widehat{a}_{log}(w_i)$, respectively. (For expositional convenience, we suppress the dependence of $\widehat{a}_{lin}$ and $\widehat{a}_{log}$ on w_A and w_B.)

Figure 1 illustrates the predicted effort using the two models described above and contrasts it to the actual effort, assuming that the principal's A/B test comprises w_2 and w_4, and the coefficient of RRA $\rho = 0.3$. The logarithmic model predicts the effort corresponding to all treatments with good accuracy— the absolute percentage error (APE), defined as

$$APE(\widehat{a}_k) := \frac{|\widehat{a}_k(w_i) - a(w_i)|}{a(w_i)} \,,$$

where $k \in \{lin, log\}$ and $a(w_i)$ denotes the actual effort, is less than 2.5% in all cases. (Except for w_1 for which it cannot make any prediction absent additional assumptions, as discussed in footnote 15.) On the other hand, the linear model predicts only the effort corresponding to w_3, w_6, and w_7 with reasonable accuracy—the APE is less than 6% for these cases, but the model grossly overestimates the effort corresponding to w_1 and w_5. By zooming in the data used to generate this figure, we see that these two treatments involve the largest change

[14] Per our assumption in this section, $v(\omega) = v_0 + \omega^{1-\rho}/(1-\rho)$ for some constant v_0. The desired expression follows from the fact that $\int f_a(x|a)dx = 0$ for any a.

[15] The first treatment in Table 2, w_1, yields $M(w_1) = 0$ (since $\int f(x|a)dx = 0$ for any a), so that ϵ, and hence (15) is not defined. This observation, together with the fact that $a(w_1) = 1521 > 0$, suggests that subjects may also be motivated by factors other than direct monetary compensation, such as the prospect of a good M-Turk rating, or they might be intrinsically motivated. A simple way to incorporate such indirect incentives is to assume that each agent's marginal benefit of effort is $M(w) + I$, where I is a parameter to be estimated and is meant to capture such indirect incentives. See Appendix A.5 for details. In light of this issue, we do not consider the logarithmic model for pairs that include w_1.

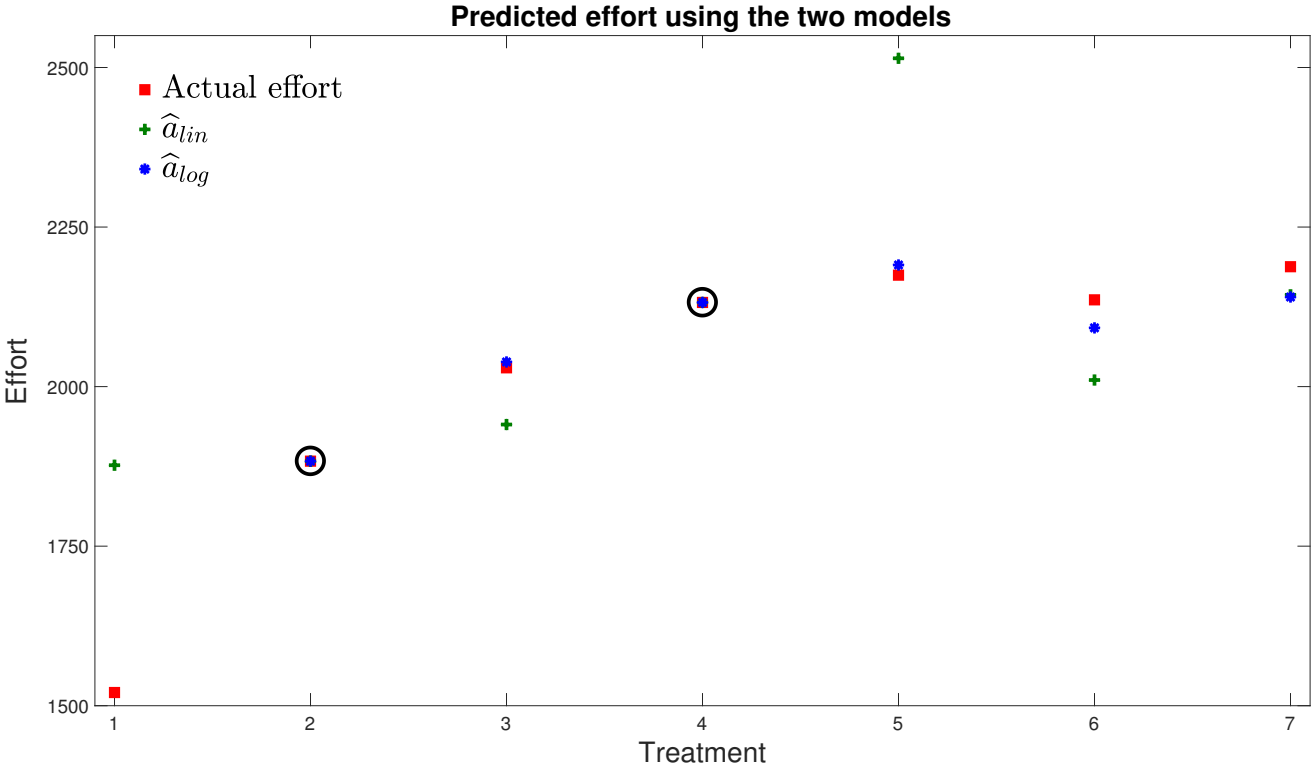

Figure 1: Predicted versus actual effort assuming that the principal has data for treatments 2 and 4, and the coefficient of RRA $\rho = 0.3$.

in the marginal incentives relative to the A/B test. Unsurprisingly, the linear model cannot make accurate predictions far out-of-sample.

Let us define for each available A/B test, the mean and the worst-case APE as

$$\text{Mean APE}(\widehat{a}_k) = \frac{1}{4}\sum_{i\notin\{A,B\}}\left|\frac{\widehat{a}_k(w_i) - a(w_i)}{a(w_i)}\right| \quad \text{and} \quad \text{Max. APE}(\widehat{a}_k) = \max_{i\notin\{A,B\}}\left|\frac{\widehat{a}_k(w_i) - a(w_i)}{a(w_i)}\right|,$$

respectively. Figure 3 presents the mean and the worst-case APE for every available A/B test, which the principal uses to predict the effort corresponding to the other treatments. (There are 15 A/B tests available. For the reason described above, we have ignored A/B tests that involve w_1 from this figure.) For example, treatment pair 3-6 refers to the case in which the principal's A/B test comprises w_3 and w_6. The average APE across all pairs is 2.08% and 6.06% for the logarithmic and the linear model, respectively. The logarithmic model outperforms the linear model in every case. This is not surprising, considering that we are predicting out of sample, and the estimated $\epsilon < 0.05$ (from (15)) in all A/B tests, while the linear model implicitly assumes unit elasticity. In fact, in all but three cases, the mean and worst-case APE for the logarithmic model is less than 2% and 3.4%, respectively. For

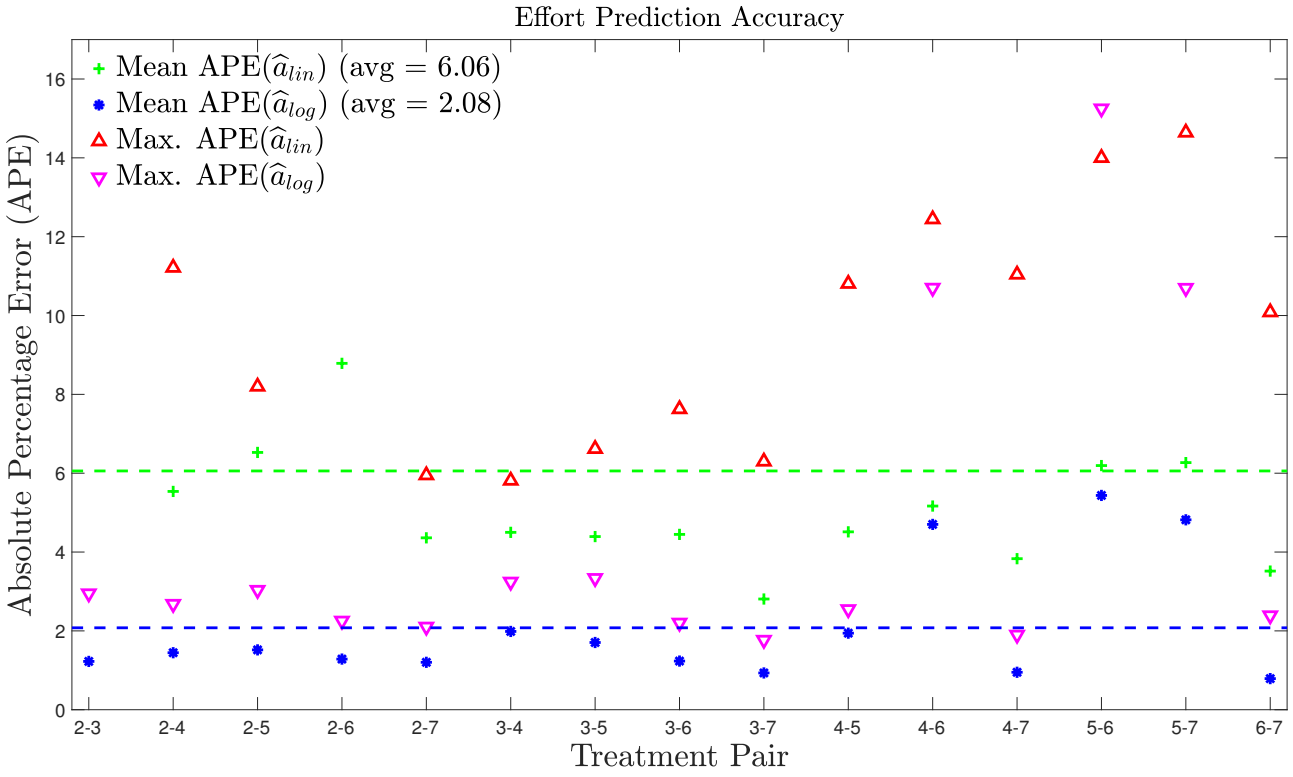

Figure 2: Effort prediction accuracy (coefficient of RRA $\rho = 0.3$).

the purpose of comparing these values to the dispersion in the data, the average (median) *absolute error* across all predictions is 125.3 (93.2) and 42.3 (31.1) for the linear and the logarithmic model, respectively, while the standard deviation of effort is 115.2.

Recall that the principal must take a stance on the agent's marginal utility function, and in this section, we have assumed that subjects' utility exhibits constant RRA. To evaluate how each model's prediction accuracy depends on this assumption, first, we varied the coefficient of RRA, ρ, from zero to one. Next, we considered the assumption that subjects have quadratic utility, and so $v'(\omega) = A - 2B\omega$, where we normalized $A = 10^3$ and we varied $B \in [10^{-3}, 1]$.[16] Figure 3 illustrates the average APE (across the 15 treatment pairs). Observe that in both cases, the logarithmic model outperforms the linear model. Interestingly, the prediction accuracy of the former is essentially invariant to whether the principal assumes that the agent has CRRA or quadratic utility, as well as to the assumed coefficient of risk aversion.

Throughout this paper, we take the A/B test that the principal has at her disposal as exogenous, and seek how to best exploit it. In practice of course, what test to conduct is itself a choice, and the principal may want to choose one that enables her to make more accurate predictions. While analyzing this problem is beyond the scope of this paper, a deeper look

[16] We chose this range for B to ensure that the marginal utility $v'(w_i(x)) > 0$ for all i and x.

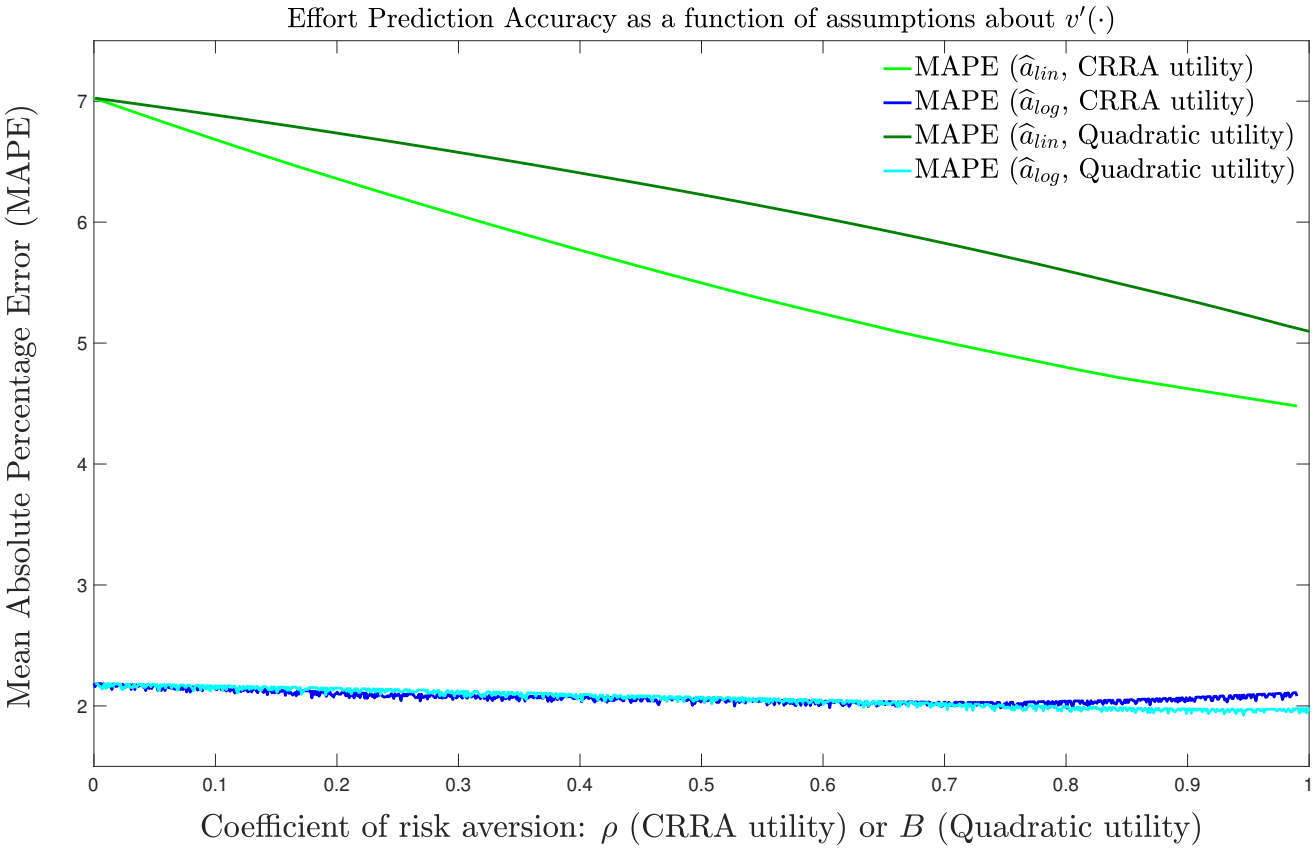

Figure 3: Effort prediction accuracy as a function of the principal's assumption about the agent's marginal utility function v'

at the data behind Figure 2 can shed light on what distinguishes A/B tests that enable more accurate predictions. Let us focus on the nonlinear model, and observe that the A/B tests comprising $\{w_4, w_6\}$, $\{w_5, w_6\}$, and $\{w_5, w_7\}$ generate poorer predictions compared to the other tests. To see why, let us consider the test comprising $\{w_4, w_6\}$. Observe from Table 1 that w_4 and w_6 lead to nearly identical effort, and from Figure 5 that the corresponding pdf's are markedly different. This is because under w_6, subjects receive a lump-sum bonus if they exceed 2,000 keystrokes, and so they reduce the intensity of their efforts once they exceed that threshold. Thus, in computing $\widehat{f}_a$ using (2), the denominator is close to zero, while the numerator is not. The same problem arises when the A/B test comprises $\{w_5, w_6\}$ or $\{w_5, w_7\}$. This problem does not arise when the A/B test comprises any of the other affine contracts and w_6 or w_7, because the difference between the corresponding efforts, and hence the denominator of (2), is sufficiently far from zero. This problem also does not arise when the A/B test comprises $\{w_6, w_7\}$. Because subjects slow down once they exceed 2,000 keystrokes under both contracts, the numerator of (2) is not too far away from zero.

Next, we turn to the principal's profit. Using either the linear model, or the logarithmic

model to predict effort, and using $(\widehat{P})$, we obtain the following prediction for the principal's profit if the status quo contract, w_A, is replaced by w:

$$\widehat{\pi}_k(w) = m\widehat{a}_k(w) - \int w\left(\widehat{f} + [\widehat{a}_k(w_j) - a(w_A)]\,\widehat{f}_a\right) dx\,, \tag{17}$$

where $k \in \{lin, log\}$. Figure 4 illustrates the mean and the worst-case APE for every pair of available A/B tests. A similar pattern to Figure 2 emerges: The logarithmic model substantially outperforms the linear model, and its mean APE is below 2.1% except for the A/B tests comprising treatments $\{4,6\}$, $\{5,6\}$, and $\{5,7\}$. For the purpose of comparing these values to the dispersion in the data, the average (median) absolute error across all predictions is 13.2 (10.7) and 5.4 (3.7) for the linear and the logarithmic model, respectively, while the standard deviation of profit is 68.8. We conclude this section with the following

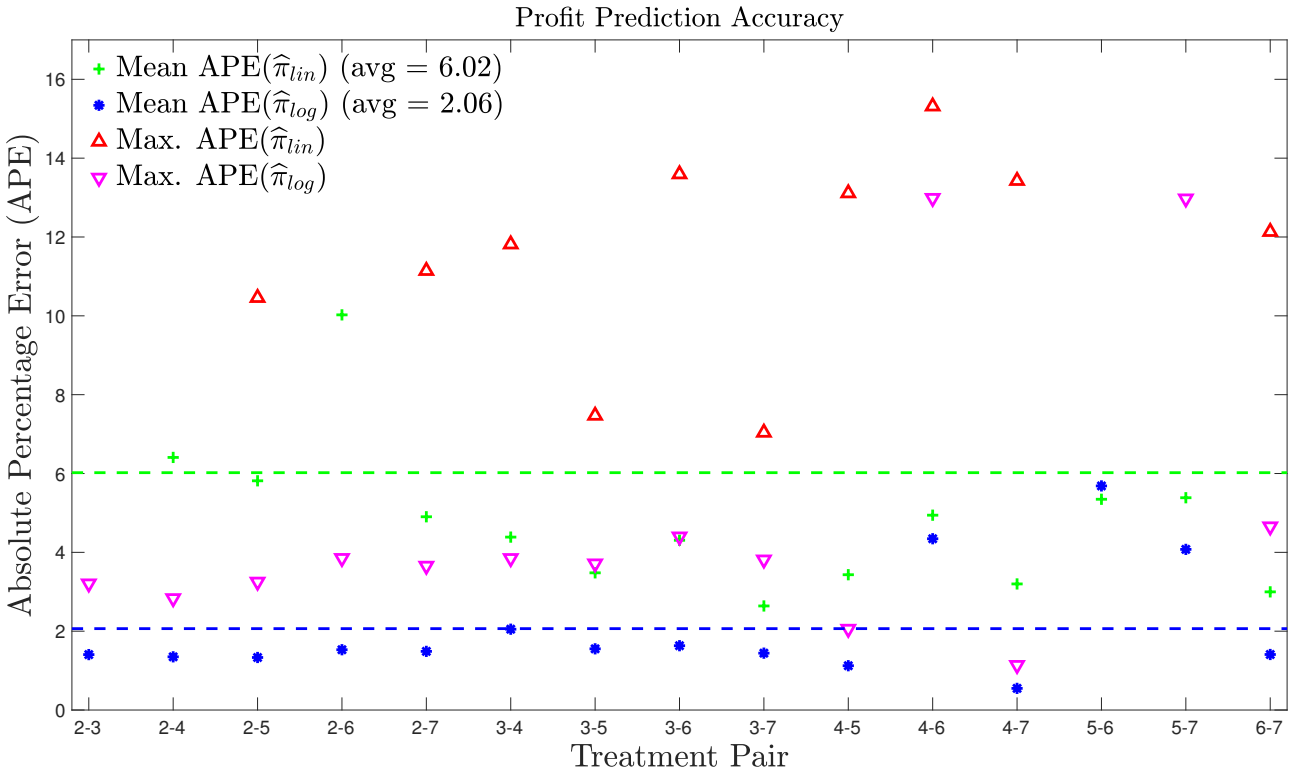

Figure 4: Profit prediction accuracy ($m = 0.2$ and coefficient of RRA $\rho = 0.3$).

remark: If one pursues to predict the profit corresponding to an affine contract (*e.g.*, for w_j where $j \in \{1,..,5\}$), then the prediction accuracy is equal to the prediction accuracy for effort. For example, if w_j has slope α and we evaluate prediction accuracy by the APE, then

$$\text{APE}(\widehat{\pi}_k) = \left|\frac{\widehat{\pi}_k(w_j) - \pi(w_j)}{\pi(w_j)}\right| = \left|\frac{(m-\alpha)\widehat{a}_k(w_j) - (m-\alpha)a(w_j)}{(m-\alpha)a(w_j)}\right| = \text{APE}(\widehat{a}_k)\,,$$

where $\pi(w_j)$ denotes the "actual" profit, computed using the experimental data. However, to predict the profit associated with a nonlinear contract such as w_6 and w_7, one must predict not only the corresponding change in effort, but also the change in the entire distribution of output. Thus, it is not surprising that the prediction accuracy for profits is not as good as the prediction accuracy for effort (when evaluated by the worst case APE), as can be seen in Figures 3 and 4.

5.2 Counterfactuals

The goal of this section is to illustrate an application of the methodology described in Section 4, and evaluate its performance. To do so, we first posit a model for the agent's problem, and we use the data corresponding to all seven treatments given in Table 1 to estimate its parameters; *i.e.,* the agent's utility function, his cost function, and the family of pdf's that map effort into output.

Next, we pick an arbitrary pair of treatments, denoted w_A and w_B, to constitute our A/B test. To obtain a benchmark, we use the estimated model to find the optimal contract that gives the agent at least as much expected utility as w_A; *i.e.*, we solve the problem posed in Section 2.1, where $\underline{u}$ is set equal to the agent's expected utility when he is offered w_A. Then we characterize the optimally perturbed contract by first solving $(\widehat{P})$ for all Δa, and then solving (14) to find the profit-maximizing Δa. Finally, we use the estimated model to compute the corresponding projected profit, and we compare it to that of the benchmark and the status quo contract, w_A.

Step 1: Estimate the Model

We begin by estimating $f(x|a)$ and $f_a(x|a)$ for all x and a in the relevant range. We will assume that x is a continuous random variable and takes values in $[0, 3500]$. Letting a_i denote the effort corresponding to treatment i given in Table 1, we use the triweight kernel to estimate $f^i(x)$ for each $i \in \{2, 3, 4, 5, 7\}$ (Hansen, 2009).[17] Then, we define for every x,

$$f_a^i(x) := \frac{f^i(x) - f^{i-1}(x)}{a_i - a_{i-1}}\,.$$

[17]Because $a_4 \simeq a_6$, we ignore treatment 6 in this step. Thus, for $i = 7$, we abuse notation and write a_{i-1} to imply a_5.

Letting $\theta_i := (a_i + a_{i-1})/2$ for each $i \in \{2, 3, 4, 5, 7\}$, we assume that

$$f_a(\cdot|a) \equiv f_a^2(\cdot) \quad \text{if } a \leq \theta_2\,,$$
$$f_a(\cdot|a) \equiv \frac{a-\theta_i}{\theta_{i+1}-\theta_i} f_a^{i+1}(\cdot) + \frac{\theta_{i+1}-a}{\theta_{i+1}-\theta_i} f_a^i(\cdot) \quad \text{if } a \in [\theta_i, \theta_{i+1}]\ \text{, and}$$
$$f_a(\cdot|a) \equiv f_a^7(\cdot) \quad \text{if } a \geq \theta_7.$$

Now we define $f(\cdot|a_1) \equiv f^1(\cdot)$, and recursively, for all $a \in (a_1, 2200]$, $f(\cdot|a) \equiv f(\cdot|a_1) + \int_{a_1}^{a} f_a(\cdot|s)ds$.[18] Figure 5 illustrates the empirical cumulative distribution function and $f(\cdot|a)$

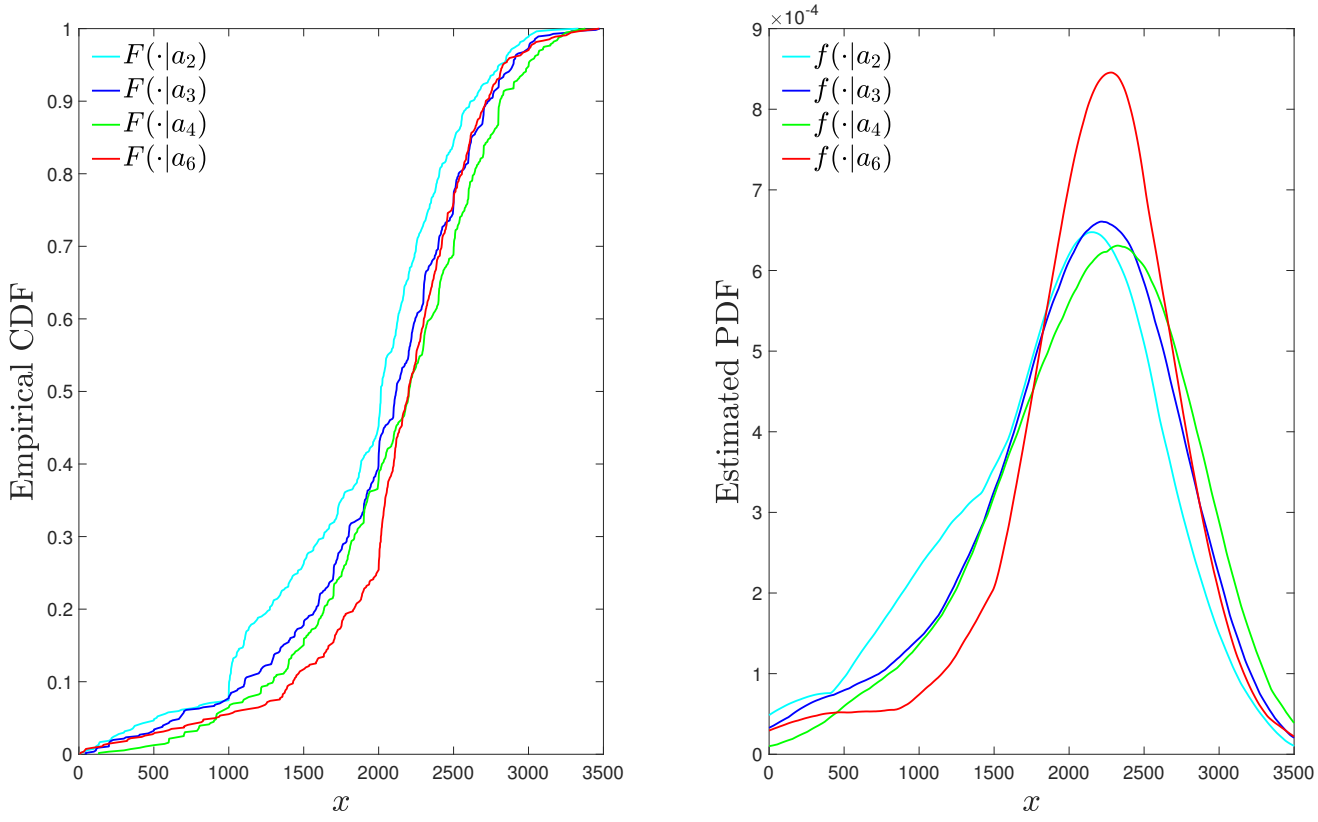

Figure 5: Empirical CDF and estimated pdf (using the triweight kernel) corresponding to treatments 2, 3, 4 and 6

agent's model, we assume that he has utility function $v(\omega) = \omega^{1-\rho}/(1-\rho)$ and cost function $c(a) = c_0 a^{p+1}/(p+1)$, for some parameters $\rho \in [0, 1)$, $c_0 > 0$ and $p > 0$. To rationalize the fact that subjects exert strictly positive effort even in the first treatment (where they are not offered any explicit monetary incentives), we assume that given contract w, the agent

[18] For $a > 2200$, the above algorithm yields $f(x|a) < 0$ for some x, which violates the definition of a pdf; hence, we restrict attention to $a \in [a_1, 2200]$.

ρ	c_0	p	I
0.3	5.797×10^{-97}	28.286	6.163×10^{-7}

Table 2: Estimates for the unknown parameters of the agent's problem

chooses a such that

$$\int v(w(x)) f_a(x|a) dx + I = c'(a)\,, \tag{18}$$

where I is a parameter that captures indirect incentives or intrinsic motivation.[19] Table 2 reports the estimates for unknown parameters using nonlinear least squares minimization. Moreover, we assume that the principal's marginal profit, $m = 0.2$.

Step 2(a): Optimal Contract (Benchmark)

We now pick an arbitrary pair of treatments (excluding treatment 1 for the reasons explained in footnote 15) to form the principal's A/B test, which we denote by w_A and w_B, respectively. We denote the principal's profit corresponding to w_A by Π_A.

To obtain our benchmark, we compute $\Pi(a)$ for every $a \in [a_1, 2200]$ by solving (P_H), where we use the estimated parameters given in Table 2 and we set $\underline{u} := \int v(w_A) f(\cdot|a(w_A)) dx - c(a(w_A))$. We incorporate two additional constraints into the problem: first, that $w(x) \geq 100$ to capture that each subject must be paid a participation fee of 100 cents, and second, that $w(x)$ is weakly increasing in x. This assumption aims to make the contract more realistic, as it is unlikely that a manager would implement a non-increasing contract. Then, using line-search on a, we find the optimal contract that gives the agent at least as much utility as w_A, and the corresponding profit, denoted w^* and Π^*, respectively.[20]

Step 2(b): Optimal Perturbation

Given an A/B test, we estimate $f(\cdot|a(w_i))$ for $i \in \{A, B\}$, denote $\widehat{f}(\cdot) \equiv f(\cdot|a(w_A))$, and define

$$\widehat{f}_a(\cdot) \equiv \frac{f(\cdot|a(w_B)) - f(\cdot|a(w_A))}{a(w_B) - a(w_A)}\,.$$

Next, we must make an assumption about the principal's stance on the agent's marginal utility function. We assume that she (correctly) guesses that $v'(\omega) = \omega^{-\rho}$ with $\rho = 0.3$; *i.e.*, that the agent has isoelastic utility with coefficient of RRA equal to 0.3. (We consider alternative assumptions at the end of this section.)

[19] See Appendix A.4 for a formal treatment of this case.

[20] Note that w^* and Π^* depend on the underlying A/B test. For expositional convenience, we suppress this dependence.

Given these parameters, we solve $(\widehat{P})$ subject to $(\widehat{IC})$, $(\widehat{IR})$, and the two additional constraints (*i.e.*, that $w \geq 100$ and it is monotonically increasing) for every Δa such that $\widehat{f}(x) + \Delta a \widehat{f}_a(x) \geq 0$ for all x. Then we find the Δa that is predicted to lead to the largest profit, and we denote the corresponding by $\widehat{w}^*$.[21] To evaluate its profit, denoted $\widehat{\Pi}^*$, we use the estimated model from Step 1.

We also solve for every Δa, $(\widehat{P}_{lin})$ subject to the two additional constraints, and the constraint that $\|t\|_1 \leq 100$. Because that problem is linear, we bound the "length" of a perturbation to ensure that an optimal solution exists. Then we find the Δa that is predicted to lead to the largest profit, and we denote the corresponding contract and profit by $\widehat{w}^*_{lin}$ and $\widehat{\Pi}^*_{lin}$, respectively.

Step 3: Evaluation

We now evaluate the performance of our methodology. In particular, we are interested in determining the extent to which the optimally perturbed contract, $\widehat{w}^*$, increases the principal's profits relative to the status quo contract, w_A, and how close it can get (profit-wise) to the optimal contract, w^*.

Figure 6 illustrates, for the case in which the principal's A/B test comprises treatments 4 (w_A) and 7 (w_B), the status quo contract, the benchmark contracts, as well as the optimally perturbed contract, and reports the corresponding profits. The contract $\widehat{w}^i_{costmin}$, $i \in \{A, B\}$ is the solution of $(\widehat{P})$ for $\Delta a \in \{0, a(w_B) - a(w_A)\}$. This corresponds to the profit-maximizing contract (based on the data available by the A/B test) that gives the agent at least as much utility as w_A and induces him to choose $a(w_A)$ and $a(w_B)$, respectively. Naturally, the optimal contract, w^*, generates the largest profit. The optimally perturbed contract, $\widehat{w}^*$ generates slightly smaller profit, but still, it achieves over 90% of the profit gap between w_A and the optimal contract. While $\widehat{w}^B_{costmin}$ achieves similar profit (and the contract itself is similar to $\widehat{w}^*$), perhaps surprisingly, $\widehat{w}^A_{costmin}$ performs poorly—worse than w_A. This is not uncommon in this dataset, and it appears to be due to the noise associated with the estimated $\widehat{f}$ and $\widehat{f}_a$.

Figure 7 illustrates how much of the profit gap between the status quo contract and the optimal contract, the optimally perturbed contract captures. To be specific, it illustrates, for each available A/B test, as a percentage of the profit of the optimal contract, (I) the profit corresponding to the status quo contract, (II) the profit of the best cost-minimizing contract as estimated using the data contained in her A/B test, and (III) the profit corresponding to the optimally perturbed contract that solves $(\widehat{P})$ and $(\widehat{P}_{lin})$, using blue upward-pointing

[21] We remark that we choose the profit-maximizing Δa using the data contained in the A/B test (as opposed to the estimated model from step 1).

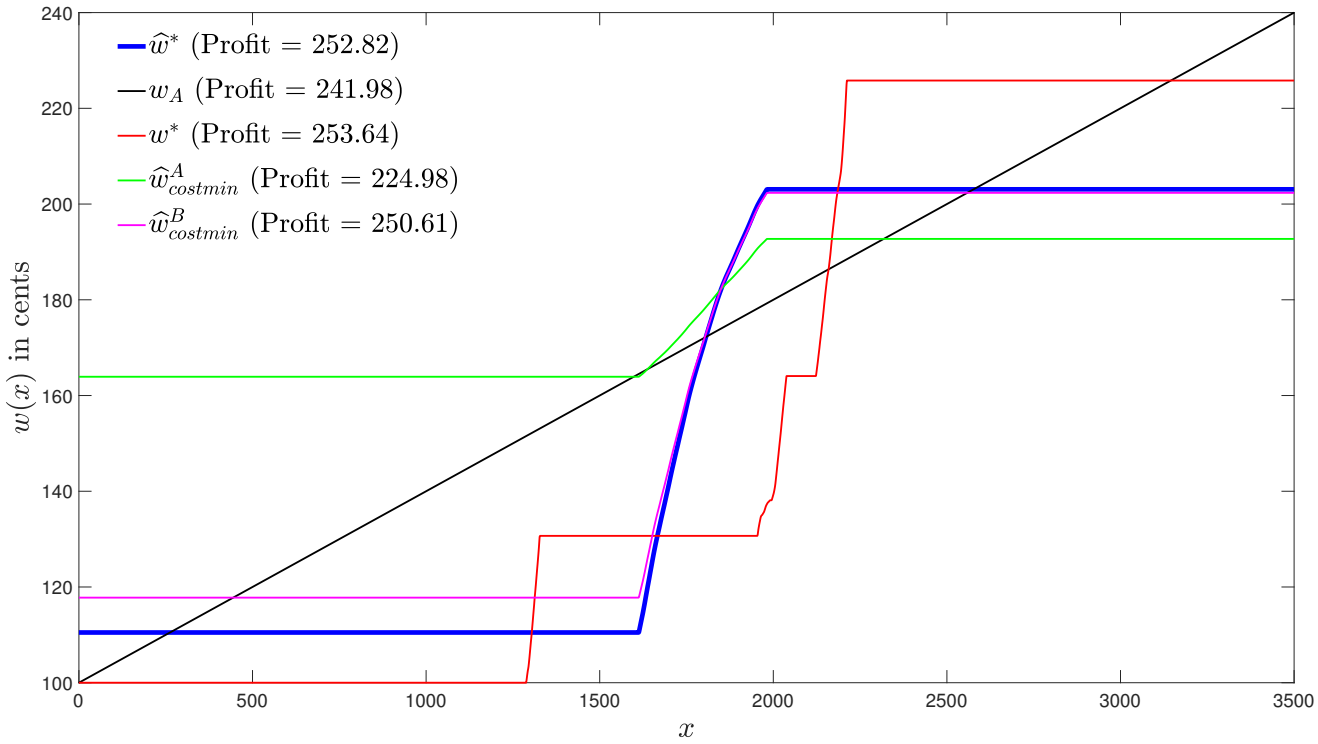

Figure 6: Optimally perturbed contract and benchmarks for the case in which the A/B test comprises treatments 4 and 7 ($m = 0.2$ and coefficient of RRA $\rho = 0.3$)

triangles and red pentagrams, respectively.[22]

Averaging across all A/B tests, the optimally perturbed contract that solves $(\widehat{P})$ and $(\widehat{P}_{lin})$ achieves 97.04% and 97.25% of the profits of the corresponding optimal contract, respectively. Meanwhile, the status quo contract and the best cost-minimizing contract achieves 94.72% and 95.24% of the profits of the optimal contract, respectively.

Recall that our goal is to design a perturbed contract that increases the principal's profits and gives the agent at least as much expected utility as the status quo contract w_A. Inspection of the agent's expected utility under the optimal perturbation that solves $(\widehat{P})$, $\widehat{w}^*$, yields the following observations: First, $\widehat{w}^*$ gives, on average, 0.8% more utils to the agent compared to the respective w_A. Second, in all but four cases, $\widehat{w}^*$ gives the agent at least 99% of the utility that w_A does. Those worst cases correspond to treatment pairs "3-4", "4-6", "4-7", and "5-7", where $\widehat{w}^*$ gives the agent 98.6%, 96.1%, 97.2%, and 97.8% of the utility that w_A does, respectively.

[22] As an example, treatment pair "4-7" refers to the case in which w_A and w_B corresponds to treatment 4 and 7, respectively. Note that pair "4-7" differs from "7-4" in that in the former (latter) case, the principal is looking for contracts that give the agent at least as much utility as w_4 (w_7), respectively. Note that for the treatment pairs "5-6" and "6-5", our (logarithmic) model is unable to generate a prediction, because it estimates $\epsilon < 0$, which is problematic.

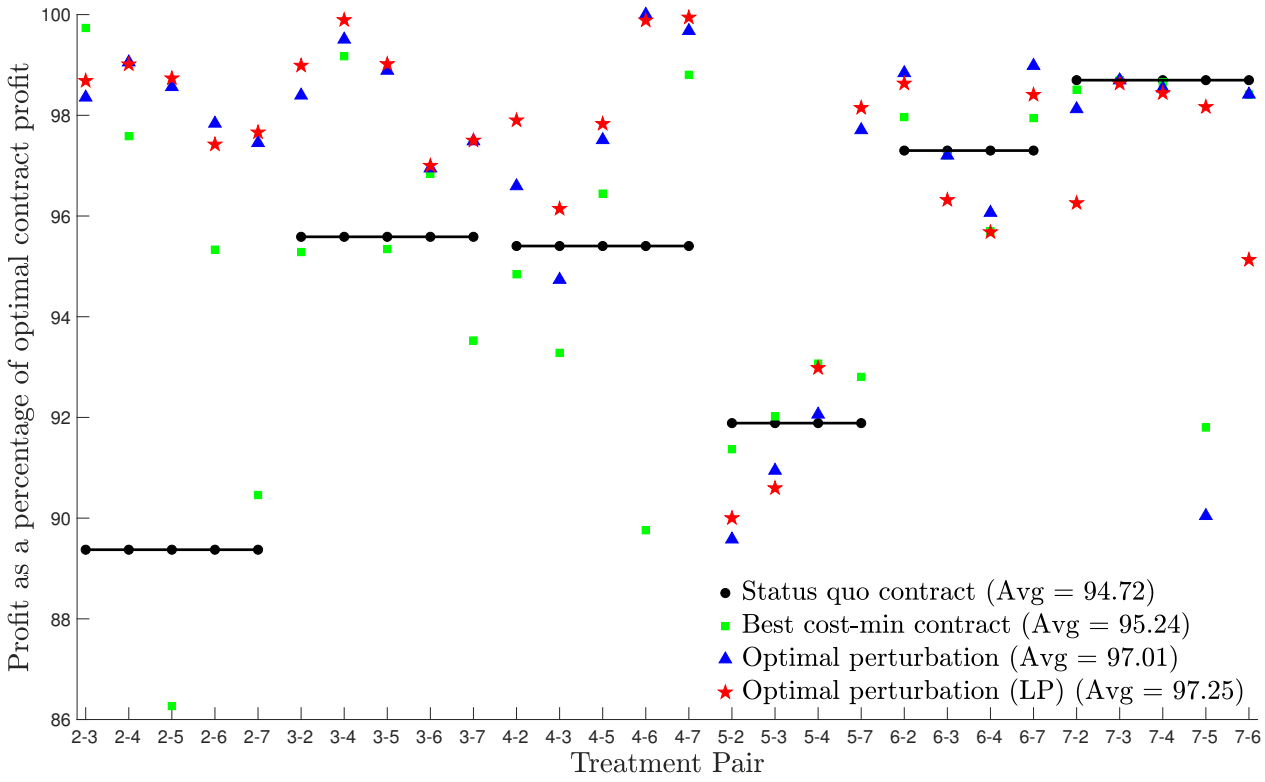

Figure 7: Profits of optimally perturbed contract relative to optimal contract and other benchmarks for every A/B test ($m = 0.2$ and coefficient of RRA $\rho = 0.3$)

The fact that the best cost-minimizing contract only marginally improves upon w_A, while the optimal perturbation brings a more substantial profit increase suggests that in perturbing the status quo contract, the profit gains lie primarily in finding the profit-maximizing effort rather than the contract that induces the agent to choose a particular effort level at minimum cost.

6 Discussion

We consider the problem faced by a firm who wants to optimize the performance pay plan that she offers to her employee(s). Using a canonical principal-agent framework a-la-Holmström (1979), we begin with the premise that she has productivity data corresponding to two different contracts (*e.g.,* an A/B test), and seeks a new contract that increases profits by as much as possible. We show that this data is a sufficient statistic for the question of how best to locally improve a status quo incentive contract, given a priori knowledge of the agent's monetary preferences. We assess the empirical relevance of this result using a dataset from DellaVigna and Pope (2017). Finally, we discuss how our framework can be extended to incorporate additional considerations beyond those in the classic theory.

References

Baker, G., 2000. The use of performance measures in incentive contracting. *American Economic Review*, 90(2), pp.415-420.

Balbuzanov, I., Gars, J. and Tjernstrom, E., 2017. Media and Motivation: The Effect of Performance Pay on Writers and Content. Working Paper.

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

Bandiera, O., Barankay, I. and Rasul, I., 2009. Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica*, 77(4), pp.1047-1094.

Barron, D., Georgiadis, G. and Swinkels, J., 2019. Optimal Contracts with a Risk-Taking Agent. Working Paper.

Barseghyan, L., Molinari, F., O'Donoghue, T. and Teitelbaum, J.C., 2018. Estimating risk preferences in the field. *Journal of Economic Literature*, 56(2), pp.501-64.

Bolton, P. and Dewatripont, M., 2005. Contract Theory. MIT Press.

Boyd, S. and Vandenberghe, L., 2004. Convex optimization. Cambridge university press.

Brewer, M., Saez, E. and Shephard, A., 2010. Means-testing and Tax Rates on Earnings. In *Dimensions of Tax Design: The Mirrlees Review*.

Carroll, G., 2015. Robustness and linear contracts. *American Economic Review*, 105(2), pp. 536-563.

Chade, H. and Swinkels, J., 2019. Disentangling moral hazard and adverse selection. Working Paper.

Chiappori, P.A. and Salanié, B. 2003. Testing Contract Theory: A Survey of Some Recent Work, in M. Dewatripont, L. Hansen and S. Turnovsky, eds., Advances in Economics and Econometrics, Cambridge University Press, Cambridge.

DellaVigna, S. and Pope, D., 2017. What motivates effort? Evidence and expert forecasts. *Review of Economic Studies*, 85(2), pp.1029-1069.

Diamond, P.A., 1998. Optimal Income Taxation: An Example with a U-shaped Pattern of Optimal Marginal Tax Rates. *American Economic Review*, pp.83-95.

Fehr, E. and Goette, L., 2007. Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review*, 97(1), pp.298-317.

Gayle, G.L. and Miller, R.A., 2009. Has moral hazard become a more important factor in managerial compensation?. *American Economic Review*, 99(5), pp.1740-69.

Garrett, D.F. and Pavan, A., 2015. Dynamic Managerial Compensation: A Variational Approach. *Journal of Economic Theory*, 159, pp.775-818.

Georgiadis, G. and Szentes, B., 2019. Optimal Monitoring Design. Working Paper.

Gibbs, M., 2016. Past, present and future compensation research: Economist perspectives. *Compensation & Benefits Review*, 48(1-2), pp.3-16.

Gibbs, M., Neckermann, S. and Siemroth, C., 2017. A field experiment in motivating employee ideas. *Review of Economics and Statistics*, 99(4), pp.577-590.

Grant, M. and Boyd, S. 2013. CVX: Matlab software for disciplined convex programming. http://cvxr.com/cvx.

Grossman, S. and Hart, O.D., 1983. An Analysis of the Principal-Agent Problem. *Econometrica,* 51(1), pp.7-45.

Guiteras, R.P. and Jack, B.K., 2018. Productivity in piece-rate labor markets: Evidence from rural Malawi. *Journal of Development Economics*, 131, pp.42-61.

Hall, B.J., Lazear, E., and Madigan, C., 2000. Performance Pay at Safelite Auto Glass (A). Harvard Business School Case 800-291.

Hansen, B.E., 2009. Lecture notes on nonparametrics.

Herweg, F., Müller, D. and Weinschenk, P., 2010. Binary payment schemes: Moral hazard and loss aversion. *American Economic Review*, 100(5), pp.2451-77.

Hill, A.E., 2019. The Minimum Wage and Productivity: A Case Study of California Strawberry Pickers. Working Paper.

Holmström, B., 1979. Moral Hazard and Observability. *Bell Journal of Economics*, pp.74-91.

Holmström, B., 2017. Pay for performance and beyond. *American Economic Review*, 107(7), pp.1753-77.

Holmström, B. and Milgrom, P., 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*, pp. 303-328.

Holmström, B. and Milgrom, P., 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, 7, pp.24-52.

Hong, F., Hossain, T., List, J.A. and Tanaka, M., 2018. Testing the theory of multitasking: Evidence from a natural field experiment in Chinese factories. *International Economic Review*, 59(2), pp.511-536.

Innes, R.D., 1990. Limited liability and incentive contracting with ex-ante action choices. *Journal of Economic Theory*, 52(1), pp.45-67.

Jensen, M.C., 2001. Corporate Budgeting Is Broken, Let's Fix It. *Harvard Business Review*, 79(10), pp.94-101.

Jewitt, I., 1988. Justifying the First-order Approach to Principal-agent Problems. *Econometrica*, pp. 1177-1190.

Jewitt, I., Kadan, O. and Swinkels, J.M., 2008. Moral Hazard with Bounded Payments. *Journal of Economic Theory*, 143(1), pp.59-82.

Koller, F., 2010. Spark: How Old-Fashioned Values Drive a Twenty-First-Century Corporation: Lessons from Lincoln Electric's Unique Guaranteed Employment Program. PublicAffairs.

Lazear, E.P., 2000. Performance pay and productivity. *American Economic Review*, 90(5), pp.1346-1361.

Mirrlees, J.A., 1976. The Optimal Structure of Incentives and Authority within an Organization. *Bell Journal of Economics*, pp. 105-131.

Oettinger, G.S., 2001. Do piece rates influence effort choices? Evidence from stadium vendors. *Economics Letters*, 73(1), pp.117-123.

Oyer, P. and Schaefer, S., 2011. Personnel Economics: Hiring and Incentives. Volume 4, Part B, Chapter 20 of Handbook of Labor Economics.

Paarsch, H.J. and Shearer, B.S., 1999. The response of worker effort to piece rates: evidence from the British Columbia tree-planting industry. *Journal of Human Resources*, pp.643-667.

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

Prendergast, C., 2014. The empirical content of pay-for-performance. *Journal of Law, Economics, & Organization*, 31(2), pp.242-261.

Saez, E., 2001. Using Elasticities to derive Optimal Income Tax Rates. *Review of Economic Studies*, 68(1), pp.205-229.

Shearer, B., 2004. Piece rates, fixed wages and incentives: Evidence from a field experiment. *Review of Economic Studies*, 71(2), pp.513-534.

Tirole, J., 1988. The Theory of Industrial Organization. MIT Press.

Wilson, R.B., 1993. Nonlinear pricing. Oxford University Press.

A Extensions

In this section, we consider four extensions of the algorithm presented in Section 4. First, we consider the case in which the principal offers a contract to a group of heterogeneous agents. Second, we incorporate multitasking by considering multidimensional effort. Next, we consider the case in which the principal restricts attention to a particular parametric class of contracts. Finally, we consider the possibility that the agent is also motivated by things besides direct financial incentives, such as the prospect of a promotion, prestige, the threat of firing, or is intrinsically motivated.

A.1 Heterogeneous Abilities

So far, we have assumed that the principal contracts with one or more homogeneous workers. In reality of course, a firm's workforce comprises heterogeneous workers, and due to practical considerations, firms often offer the same contract to all workers in a given job. The goal of this section is to extend the algorithm developed in Section 4 to the case in which the firm offers a common contract to a group of agents with heterogeneous effort costs.

Let $\theta \in \mathbb{N}$ denote the type of each agent. We assume that agents with different types have different costs of effort but are otherwise identical. If an type-θ agent, chooses effort a, then output $x \sim f(\cdot|a)$ and $\mathbb{E}[x|a] = a$. To be consistent with the analysis in Section 4, we maintain Assumptions 2-3, thus assuming that $f_a(\cdot|a)$ does not depend on a, and the cost of a type-θ agent choosing effort a is equal to $c^\theta(a) = c_0^\theta a^{1/\epsilon^\theta}$ for some constants c_0^θ and ϵ^θ. Let $a^\theta(w)$ denote the effort chosen by a type-θ agent when offered contract w. We assume that the principal has output data corresponding to two contracts, w_A and w_B, and she knows $\widehat{f}^\theta := f(\cdot|a^\theta(w_A))$ for every θ, and $\widehat{f}_a := f_a(\cdot)$. Finally, we denote the fraction of type-θ agents by p^θ.

Since $\widehat{f}_a$ is invariant in a and θ by assumption, the marginal incentive of effort corresponding to any contract, $M(w) = \int v(w)\widehat{f}_a dx$, does not depend on a or θ. Therefore, for each θ, the principal can estimate the constants c_0^θ and ϵ^θ by solving the following system of linear equations:

$$\log a^\theta(w_i) = \beta^\theta + \epsilon^\theta \log M(w_i) \quad \text{for } i \in \{A, B\}\,. \tag{19}$$

We will argue that under the assumptions described above, the principal's problem is

expressed by the following optimization program:

$$\max_{w(\cdot),\,\Delta a^1} \ \sum_\theta p^\theta \left[m\left(\widehat{a}^\theta + \Delta a^\theta\right) - \int w\left(\widehat{f}^\theta + \Delta a^\theta \widehat{f}_a\right) dx \right] \tag{$\widehat{P}^\theta$}$$

$$\text{s.t.} \ \int v(w)\widehat{f}_a dx = \left(\frac{\widehat{a}^1 + \Delta a^1}{\widehat{a}^1}\right)^{1/\epsilon^\theta} \int v(w_A)\widehat{f}_a dx \tag{$\widehat{IC}^1$}$$

$$\Delta a^\theta = \widehat{a}^\theta \left[\left(1 + \frac{\Delta a^1}{\widehat{a}^1}\right)^{\epsilon^\theta/\epsilon^1} - 1 \right] \quad \text{for all } \theta \geq 2 \tag{$\widehat{IC}^\theta$}$$

$$\int v(w)(\widehat{f}^\theta + \Delta a^\theta \widehat{f}_a)dx \geq \int v(w_A)\widehat{f}^\theta dx + \frac{\epsilon^\theta e^{-\beta^\theta/\epsilon^\theta}}{1+\epsilon^\theta} \left[\left(\widehat{a}^\theta + \Delta a^\theta\right)^{\frac{1+\epsilon^\theta}{\epsilon^\theta}} - \left(\widehat{a}^\theta\right)^{\frac{1+\epsilon^\theta}{\epsilon^\theta}} \right] \quad \text{for all } \theta \tag{$\widehat{IR}^\theta$}$$

where $\widehat{a}^\theta := a^\theta(w_A)$. This is the counterpart of $(\widehat{P})$, $(\widehat{IC})$, and $(\widehat{IR})$. Let us explain each expression. First, for any given target effort level Δa^1, $(\widehat{IC}^1)$ follows from (19) for $\theta = 1$. Because $M(w)$ does not depend on θ, $(1 + \Delta a^1/\widehat{a}^1)^{1/\epsilon^\theta}$ is equal to a constant (and independent of θ), so for any Δa^1 and $\theta \geq 2$, $(\widehat{IC}^\theta)$ must be satisfied. Finally, $(\widehat{IR}^\theta)$ and $(\widehat{P}^\theta)$ follow the same logic as $(\widehat{IR})$ and $(\widehat{P})$, after adding the appropriate θ-superscripts.

This program can be solved using the same two-stage approach as $(\widehat{P})$: First, for every Δa^1, one finds the contract that motivates each type-θ agent to choose effort $\widehat{a}^\theta + \Delta a^\theta$ at maximum profit. Then, one finds the profit-maximizing Δa^1 and the corresponding contract using line-search.

It is well-known that the design of performance pay may be used to induce *selection*, that is, attract more productive workers and induce less productive workers to exit (Lazear, 2000). To do so in our setting, the principal may restrict attention to perturbations that give at least as much utility as $\widehat{w}$ to only a subset of the more productive types. Formally, this would imply that $(\widehat{IR}^\theta)$ must hold only for the types that the principal wants to attract, and must be violated for the types that she wants to repel from the firm or dissuade from joining. A complete analysis of the selection effects associated with performance pay is beyond the scope of this paper, and is left for future work.

A.2 Multi-dimensional Effort *(Incomplete)*

In this section, we extend our baseline model to the case in which the agent's effort is multi-dimensional. As an example, each agent might be a salesperson, who sells different products. To be specific, suppose that for each product $i \in \{1, ..., N\}$, he chooses effort a_i, which in turn generates sales $\mathbf{x} \sim f(\cdot|\mathbf{a})$, where $\mathbf{a}$ and $\mathbf{x}$ is a (finite) vector of efforts and sales, respectively.

The cost of choosing $\mathbf{a}$ is equal to $c(\mathbf{a})$, and the function c satisfies the usual conditions. A contract w is an u.s.c mapping from sales $\mathbf{x}$ to a monetary transfer to the agent. We continue to assume that the principal's performance measure, $\mathbf{x}$, is sufficiently broad so that the agent cannot distort it (Baker, 2000).

We assume that the principal has output data corresponding to a status quo contract $\widehat{w}$, and thus she can estimate the density $\widehat{f} := f(\cdot|\mathbf{a}(\widehat{w}))$, where $\mathbf{a}(w)$ denotes the (vector of) efforts chosen by an agent when offered contract w. Given any perturbation t, which is an u.s.c mapping from $\mathbf{x}$ to a real number, let us define for each i, the Gateaux derivative

$$\mathcal{D}a_i(\widehat{w}, t) := \left.\frac{da_i(\widehat{w} + \theta t)}{d\theta}\right|_{\theta=0} = \lim_{\theta \to 0} \frac{a_i(\widehat{w} + \theta t) - a_i(\widehat{w})}{\theta}\,. \tag{20}$$

The principal's expected profit when she offers contract w,

$$\pi(w) = \sum_i m_i a_i(w) - \int w(\mathbf{x}) f(\mathbf{x}|\mathbf{a}(w)) d\mathbf{x}\,,$$

where m_i is the principal's marginal profit associated with a_i. As in Section 2, her objective is to perturb the status quo contract in the direction that increases her profit at the fastest rate; *i.e.*, solve the following maximization program

$$\max_{t \text{ u.s.c}} \mathcal{D}\pi(\widehat{w}, t) = \sum_{i=1}^{m} \left(m_i - \int \widehat{w} \widehat{f}_i d\mathbf{x} \right) \mathcal{D}a_i(\widehat{w}, t) - \int t \widehat{f} d\mathbf{x} \tag{21}$$

subject to the constraints that t gives the agent at least as much utility as $\widehat{w}$, and $\|t\|_p \leq 1$ for some $p \in \{2, 3, ..\}$. Note that f_i denotes the derivative of f with respect to a_i, and we have dropped the dependence on $\mathbf{x}$ and on the agent's effort for notational simplicity. We will show that to solve this problem, the principal must be able to estimate $\mathcal{D}a_i(\widehat{w}, \widehat{t}_k)$ for at least K perturbations, $\widehat{t}_1, .., \widehat{t}_K$, where $K \geq \lceil (N+1)/2 \rceil$.

Similar to Section 3.1, we assume that the first-order approach is valid, and so given contract w, the agent's effort satisfies for each i

$$\int v\left(w(\mathbf{x})\right) f_i(\mathbf{x}|\mathbf{a}(w))\, d\mathbf{x} = c_i(\mathbf{a}(w))\,, \tag{22}$$

where c_i denotes the derivative of c with respect to its i^{th} argument. Using (22), the constraint that any perturbation t must give the agent at least as much utility as the status quo contract can be written as

$$\int \cdots \int t v'(\widehat{w}) \widehat{f}\, d\mathbf{x} \geq 0\,.$$

Using (22), we can compute $\mathcal{D}a_i(\widehat{w},t)$ for each i in terms of primitives as

$$\underbrace{\left[c_{ii}-\int v(\widehat{w})\widehat{f}_{ii}d\mathbf{x}\right]}_{=:B_{ii}}\mathcal{D}a_i(\widehat{w},t)+\sum_{j\neq i}\underbrace{\left[c_{ij}-\int v(\widehat{w})\widehat{f}_{ij}d\mathbf{x}\right]}_{=:B_{ij}}\mathcal{D}a_j(\widehat{w},t)=\underbrace{\int tv'(\widehat{w})\widehat{f}_i d\mathbf{x}}_{=:A_i}.$$

This is the counterpart of (3) if the agent's effort is N-dimensional. To solve (21), it suffices that the principal can evaluate $\mathcal{D}a_i(\widehat{w},t)$ for every i and any perturbation t. To do so, she must (i) take a stance on the agent's marginal utility function, v', which will allow her to evaluate the vector A for any perturbation t, and (ii) estimate the matrix B. Observe that the B is symmetric, and so it contains $N(N+1)/2$ unknowns. Therefore, to be able to estimate B, the principal must have output data corresponding to at least $K \geq \lceil (N+1)/2 \rceil$ perturbations, which will then enable her to estimate $\mathcal{D}a_i(\widehat{w},\widehat{t}_k)$ for each i and k. In that case, it is straightforward to verify that (21) can be solved using the same method as (P_{local}).

A.3 Parametric Classes of Contracts

Firms sometimes restrict attention to a particular class of contracts. For instance, linear contracts are very common, as are piece-wise linear and single-bonus contracts. Such contracts may be chosen due to their simplicity, or because of considerations outside our model.[23]

In this section, we discuss how the methodology in Section 4 can be applied when the principal only considers contracts which belong to a particular parametric class, denoted by w_α, where $\alpha \in \mathbb{R}^n$ for some $n \in \mathbb{N}$ is a vector of parameters to be chosen. For example, if she restricts attention to linear contracts, then $w_\alpha(x) = \alpha_1 + \alpha_2 x$; if she restricts attention to piece-wise linear contracts with a guaranteed minimum wage, then $w_\alpha(x) = \alpha_1 + \alpha_2(x-\alpha_3)^+$; if she restricts attention to bonus contracts, then $w_\alpha(x) = \alpha_1 + \alpha_2 \mathbb{I}_{\{x \geq \alpha_3\}}$.

It is straightforward to verify that the principal's problem still solves ($\widehat{P}$) subject to ($\widehat{IC}$) and ($\widehat{IR}$), except that w is replaced by w_α, and the choice variable w is replaced by the vector α. However, this problem is generically not convex, so standard optimization techniques are not guaranteed to achieve the global maximum. When n is sufficiently small, to find the profit-maximizing α and Δa, an alternative approach is to use grid-search.

[23] For example, if gaming is a concern, then linear contracts may be optimal; see for example, Holmström and Milgrom (1987) and Barron, Georgiadis, and Swinkels (2019). If the worker is expectation-loss averse or the performance measure is endogenous, then single-bonus contracts may be optimal; see for example, Herweg et al. (2010) and Georgiadis and Szentes (2019).

A.4 Other Sources of Incentives

Workers are motivated not only by performance pay, but also by other factors, such as the prospect of a promotion, the threat of firing, prestige, and so on. To capture such indirect incentives, suppose that faced with a contract w, the agent chooses his effort $a(w)$ by solving

$$\int v(w(x))f_a(x|a(w))dx + I = c'(a(w))\,,$$

where I is a parameter to be estimated, and captures his marginal benefit from exerting effort due factors other than performance pay. To remain consistent with the analysis in Section 4, we maintain Assumptions 2-3, and additionally assume that I does not depend on effort. Then, for any contract w, the model predicts that the following relationship is satisfied:

$$\log a(w) = \beta + \epsilon \log\left[M(w) + I\right]\,,$$

where $M(w) = \int v(w)\widehat{f}_a dx$, and the parameters β, ϵ, I must be estimated from the principal's data. Notice that to pin down all three parameters, an A/B test no longer suffices—the principal needs data corresponding to (at least) three contracts.

Turning to the principal's problem, it is straightforward to verify that $(\widehat{P})$ and $(\widehat{IR})$ are unchanged when we incorporate indirect incentives. Using the above display equation, it follows that for given I, if the principal wants to induce some effort $\widehat{a}+\Delta a$, then the contract must satisfy

$$\int v(w)\widehat{f}_a dx + I = \left(\frac{\widehat{a}+\Delta a}{\widehat{a}}\right)^{1/\epsilon}\left[\int v(w_A)\widehat{f}_a dx + I\right]\,. \tag{23}$$

To find the optimally revised contract, given our assumptions, the principal solves $(\widehat{P})$ subject to $(\widehat{IR})$ and (23).

B Proofs

Proof of Proposition 1. Let $\lambda \geq 0$ and $\nu \geq 0$ denote the dual multipliers associated with the first and second constraint in (P_{local}), respectively. The Lagrangian

$$L(\lambda,\nu) = \max_t \left\{\nu + \int\left[t\left(\lambda v'(w_A)\widehat{f} + \mu v'(w_A)\widehat{f}_a - \widehat{f}\right) - \nu t^2\right]dx\right\}. \tag{24}$$

For any $\nu > 0$, we can optimize the integrand with respect to t pointwise. Noting that the integrand is differentiable with respect to t except at $t = 0$, the corresponding first-order

condition implies that

$$t_{\lambda,\nu} = \frac{\left(\lambda \widehat{f} + \mu \widehat{f_a}\right) v'(w_A) - \widehat{f}}{2\nu},$$

where t, $\widehat{f}$, $\widehat{f_a}$, and w_A are functions of x.[24]

Next, we pin down the optimal multipliers λ and ν, by turning to the dual problem, and solving the following minimization program:

$$\min_{\lambda \geq 0\,,\, \nu \geq 0} L(\lambda, \nu)\,.$$

This problem is convex, and using $t_{\lambda,\nu}$, the corresponding first-order conditions yield

$$\lambda^* = \max\left\{0, \frac{\int \left(\widehat{f} - \mu v'(w_A)\widehat{f_a}\right) v'(w_A)\widehat{f}\,dx}{\int \left(v'(w_A)\widehat{f}\right)^2 dx}\right\} \tag{25}$$

and

$$\nu^* = \frac{1}{2}\sqrt{\int \left[\left(\lambda^* \widehat{f} + \mu \widehat{f_a}\right) v'(w_A) - \widehat{f}\right]^2 dx}\,. \tag{26}$$

Thus, the optimal perturbation,

$$t^* = t_{\lambda^*,\nu^*} = \frac{\left(\lambda \widehat{f} + \mu \widehat{f_a}\right) v'(w_A) - \widehat{f}}{\sqrt{\int \left[\left(\lambda^* \widehat{f} + \mu \widehat{f_a}\right) v'(w_A) - \widehat{f}\right]^2 dx}}\,.$$

First, let us characterize the solution to (P_{local}). Recall that the dual program is convex (even if the primal is not convex), because it is the pointwise minimum of affine functions. Therefore, the multipliers λ^* and ν^* obtained in (25) and (26) are necessary and sufficient for an optimum in the dual program.

We will now show that strong duality holds. Towards this goal, let Π^* denote the optimal value of the primal program given in (P_{local}). Weak duality implies that $L(\lambda^*, \nu^*) \geq \Pi^*$. Moreover, it is straightforward to verify that $t\,(\lambda^*, \nu^*)$ is feasible for (P_{local}), and λ^* and ν^* is strictly positive if and only if the respective (primal) constraint binds. This implies that the objective of (P_{local}) evaluated at $t\,(\lambda^*, \nu^*)$ is equal to $L(\lambda^*, \nu^*)$, and it must be the case that $L(\lambda^*, \nu^*) \leq \Pi^*$. Therefore, we conclude that $L(\lambda^*, \nu^*) = \Pi^*$, which proves that the perturbation $t\,(\lambda^*, \nu^*)$ is optimal for (P_{local}).

[24] If $\nu = 0$, then the integrand of (24) is linear in t, and the first-order condition implies that $\left(\lambda \widehat{f} + \mu \widehat{f_a}\right) v'(w_A) = \widehat{f}$, and hence $L(\lambda, 0) = 0$.

To complete the proof, we show that w_A is locally optimal if and only if (10) is satisfied for all x. Clearly, w_A is locally optimal if and only if the optimal perturbation $t^* = 0$ for all x, which is true only if for some $\lambda' \geq 0$, we have $\left(\lambda' \widehat{f} + \mu \widehat{f_a}\right) v'(w_A) = \widehat{f}$ for every x. Suppose this is the case (for some $\lambda' \geq 0$). Integrating both sides with respect to x and using that $\int \widehat{f_a} dx = 0$ implies that $\lambda' = \int \widehat{f}/v'(w_A) dx$. It is straightforward to verify that $t \equiv 0$ solves (24) when $\lambda = \lambda'$, and $L(\lambda', \nu) = \nu$ for every ν. Therefore, $\min_{\nu \geq 0} L(\lambda', \nu) = 0$, and weak duality implies that the value of the primal program is bounded by 0 from above. As $t \equiv 0$ is feasible for the primal, and the objective equals 0 when $t \equiv 0$, it follows that $t \equiv 0$ is indeed the optimal perturbation. $\square$