A/B Contracts*

George Georgiadis and Michael Powell†

January 18, 2020

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

We consider a standard contractual relationship between a principal and an agent as in Holmström (1979) but with a non-standard informational assumption and principal objective.

The agent faces a **contract**, $w(\cdot)$, which is an upper-semicontinuous mapping from output to payments made from the principal to the agent. The agent chooses a privately costly, non-contractible effort level $a \geq 0$ that determines the distribution over his output, which accrues to the principal. In particular, his output, $x \in \mathbb{R}$, is realized according to some cumulative

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.

distribution function, $F(x|\,a)$, with probability density function $f(x|\,a)$, which we assume is twice continuously differentiable in a. Without loss of generality, we normalize a so that $a = E[x|\,a]$, and the agent's effort can be interpreted as his expected output.

If the agent is paid ω and chooses effort level a, 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 $c', c'' > 0$. If the agent generates output x and is paid $w(x)$, the principal's profit is $mx - w(x)$. We assume that v and m are common knowledge.

We refer to the pair of functions $P \equiv (f, c)$ as the **production environment**. The agent observes P and chooses his effort level to maximize his expected utility. We assume that the first-order approach is valid, so that the agent's optimal effort choice is fully characterized by the first-order conditions of his problem. We denote by $a(w)$ the agent's optimal effort choice under contract w, and we assume that $a(w)$ is unique for all w.

The principal does not observe P but does observe outcome data from two contracts: a **status quo contract**, which we will denote w^A, and a **test contract**, which we will denote w^B. The **outcome data for a contract** w is the distribution of output generated by an agent facing that contract, that is, $f(\cdot|\,a(w))$. We will say that a contract $\tilde{w}$ **Pareto improves** w if the expected utility of the principal and the agent are at least as high under $\tilde{w}$ as under w given the production environment P.

The principal's objective is to choose a profit-maximizing contract from a set $\mathcal{W}$ that Pareto improves a status quo contract. Our choice of the set $\mathcal{W}$ will depend on the exercise we carry out. In Section 3, it will be the set of local adjustments to the status quo contract, and in Section XXX, it will be the full set of contracts.

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 ($\mathrm{P}_H$) 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 Local Adjustments

We first ask the question of how the principal should locally adjust a status quo contract. We will show that the information revealed by a single A/B test of contracts is sufficient for the principal to solve this problem.

To carry out this exercise, we will need to be able to describe how the principal's payoff changes as we locally adjust the status quo contract w^A, and this requires an important piece of terminology and notation. Given a contract w and a function $q(w)$, define the **Gateaux differential of q in the direction** t by $Dq(w,t) \equiv \lim_{\theta\to 0}\left[q(w+\theta t) - q(w)\right]/\theta$.

We will first characterize how the agent's effort and utility change as we locally adjust the contract. The agent's problem, given contract w, is

$$u(w) = \max_a \int v(w(x))\, f(x|a)\, dx - c(a).$$

We have assumed that the first-order approach is valid, so we can characterize the agent's optimal effort choice $a(w)$ under contract w by his first-order conditions. To this end, define the agent's **marginal incentives** as $I(w,a) \equiv \int v(w(x))\, f_a(x|a)\, dx$. Optimal effort equates

marginal costs to marginal incentives and is therefore implicitly defined by the equation $c'(a(w)) = I(w, a(w))$.

The following lemma shows how the agent's utility and effort change in response to a local adjustment to w in the direction t.

Lemma 1. *Adjusting a contract w in the direction t changes the agent's utility by*

$$Du(w,t) = \int t(x)\, v'(w(x))\, f(x|\, a(w))\, dx$$

and changes his effort by

$$Da(w,t) = \frac{DI(w,t)}{c''(a(w)) - \int v(w(x))\, f_{aa}(x|\, a(w))\, dx},$$

where $DI(w,t) \equiv \int t(x)\, v'(w(x))\, f_a(x|\, a(w))\, dx$.

The first part of the lemma follows directly from the envelope theorem. The second part shows that the agent's behavioral response to adjusting the contract in the direction t depends on how the adjustment affects his marginal incentives, $DI(w,t)$, as well as on the local curvature of his problem. It also implies that $Da(w,t)/DI(w,t)$ is independent of t: How the agent responds to an adjustment to the contract depends only on how that adjustment impacts his marginal incentives. This property will be important in what follows.

We will now describe the principal's problem under the assumption that she knows the production environment. Her expected profit from offering contract w is

$$\pi(w) = ma(w) - \int w(x)\, f(x|\, a(w))\, dx.$$

As she adjusts the contract in the direction t, her profits change according to her profit differential

$$D\pi(w,t) = \left(m - \int w(x)\, f_a(x|\, a(w))\right) Da(w,t) - \int t(x)\, f(x|\, a(w))\, dx.$$

The first term describes the change in the principal's gross profits per unit of expected output times the change in the expected output, and the second term captures the change in the expected payments she will make to the agent.

We are now prepared to state the principal's problem of how best to locally Pareto improve a status quo contract w^A. Given production environment P, she wants to choose the direction t that maximizes her profit differential subject to the constraint that it weakly improves the

agent's utility. That is, she solves

$$\max_{t:||t||\leq 1} D\pi\left(w^A, t\right) \text{ subject to } Du\left(w^A, t\right) \geq 0, \qquad (Adj_{local})$$

where $||\cdot||$ is the ℓ^2 norm. Adjustments have both direction and magnitude. We constrain the magnitude of the adjustment to isolate the choice of the optimal direction.

In describing this problem, we temporarily assumed that the principal knows the production environment. We will now show that she only needs to know certain local aspects of the production environment. To do so, we will be comparing the principal's problem across different production environments, and so it will be helpful to introduce the notation $(Adj_{local} - P)$ to refer to the principal's problem (Adj_{local}) when the production environment is P. Denote the agent's effort choice, the output density function, and the score under the status quo contract by $a^A = a\left(w^A\right)$, $f^A = f\left(\cdot \middle| a^A\right)$, and $f_a^A = f_a\left(\cdot \middle| a^A\right)$ and in an abuse of notation, denote the agent's effort differential under production environment P by $Da\left(w, t \middle| P\right)$. The following lemma shows what aspects of the production environment are relevant for solving $(Adj_{local} - P)$.

Lemma 2. *Take any two production environments $P = (f, c)$ and $\tilde{P} = \left(\tilde{f}, \tilde{c}\right)$ satisfying $f^A = \tilde{f}^A$ and $f_a^A = \tilde{f}_a^A$ and $Da\left(w^A, t \middle| P\right) = Da\left(w^A, t \middle| \tilde{P}\right)$ for all t. Then t^* solves $(Adj_{local} - P)$ if and only if it solves $\left(Adj_{local} - \tilde{P}\right)$.*

Lemma 2 shows that for the problem of locally Pareto improving a status quo contract, three pieces of local information are required: the output distribution under the status quo contract, how the output distribution changes locally in effort, and how the agent responds to *every* local change to the contract.

Before we show how a local A/B test provides this information, we need to introduce a couple definitions and pieces of notation. Take the production environment as given. An **A/B test** for contracts w^A and w^B is a pair $AB\left(w^A, w^B\right) \equiv \left(f^A, f^B\right)$, where f^A is the outcome data for w^A and f^B is the outcome data for w^B. A **local A/B test** for contracts w^A and w^B is a triple $LAB\left(w^A, w^B\right) \equiv \left(f^A, f_a^A, Da\left(w^A, w^B\right)\right)$ consisting of outcome data for w^A, information about how the output distribution changes locally in effort, and the agent's effort response to a change in the direction w^B. We will say that the test contract is **informative** if $Da\left(w^A, w^B\right) \neq 0$. One way of interpreting a local A/B test is that it consists of the local properties of the output distribution that the principal can construct with outcome data for w^A and outcome data for $w^A + \theta w^B$ as $\theta \to 0$.

The following lemma shows that the information provided by a local A/B test suffices for solving (Adj_{local}).

Lemma 3. *Take any two production environments* $P = (f, c)$ *and* $\tilde{P} = \left(\tilde{f}, \tilde{c}\right)$*, a status quo contract* w^A*, and an informative test contract* w^B*. The following are equivalent:*

1. $f^A = \tilde{f}^A$, $f_a^A = \tilde{f}_a^A$ *and* $Da\left(w^A, t \middle| P\right) = Da\left(w^A, t \middle| \tilde{P}\right)$ *for all* t*.*

2. $LAB\left(w^A, w^B \middle| P\right) = LAB\left(w^A, w^B \middle| \tilde{P}\right)$.

The proof of Lemma 3 shows how the information from a local A/B test can be used to construct the relevant information for solving (Adj_{local}). In particular, the local A/B test can be used directly to compute how the agent's marginal incentives changed, $DI\left(w^A, w^B\right)$. Then, using the insight from Lemma 1 that the agent's behavioral response to a change in marginal incentives is independent of the adjustment that led to that change in marginal incentives, it follows that for any t, $Da\left(w^A, t\right) = \frac{Da\left(w^A, w^B\right)}{DI(w^A, w^B)} DI\left(w^A, t\right)$. The agent's effort differential is therefore proportional to how his marginal incentives change with t, $DI\left(w^A, t\right)$, which can also be constructed from the local A/B test.

We now return to the principal's problem, (Adj_{local}). This is a convex-optimization problem and can be solved using standard methods. Define the following function, which we will call the Holmström-Mirrlees adjustment function:

$$T(x, \lambda, \mu) = \left[\lambda v'\left(w^A(x)\right) - 1\right] f\left(x \middle| a^A\right) + \mu v'\left(w^A(x)\right) f_a\left(x \middle| a^A\right).$$

Proposition 1 characterizes the optimal local adjustment.

Proposition 1. *Let* w^A *be the status quo contract. Suppose* t^* *solves* (Adj_{local})*. There exists* $\lambda^*, \mu^* \geq 0$ *such that* $t^*(x) \propto T(x, \lambda^*, \mu^*)$*. If* w^A *is locally optimal, then* $T(x, \lambda^*, \mu^*) = 0$ *for all* x*.*

The first part of this proposition shows that optimal local adjustments are in the direction of a Holmström-Mirrlees-type contract, that is, they locally balance risk allocation with incentive provision: They shift payments from outputs where the agent has a low marginal utility of money to those where his marginal utility of money is higher. And they shift payments towards outputs that increase the agent's marginal incentives. The optimal way to balance these two considerations is determined by the coefficients λ^*, μ^*, the exact expressions for which which are given in the proof of Proposition 1 in the appendix.

The second part of this proposition echoes the optimality conditions of Holmström (1979) and serves as a consistency check. When the agent's effort choice is locally optimal, the coefficients λ^* and μ^* coincide with those in Holmström (1979). The primary contribution of Proposition 1 is to show how λ^* and μ^* change as we consider status quo contracts that are not locally optimal. In particular, μ^*, the weight that is optimally put on how marginal

incentives are adjusted, is higher when the principal's expected gains from a higher effort level are higher and when the agent's response to an increase in marginal incentives is higher. The weight that is put on the risk-allocation component, λ^*, is smaller when μ^* is higher.

4 Non-local Adjustments

The analysis in Section XXX illustrates how local information suffices for characterizing optimal local adjustments. This section provides a method for extrapolating from local information to assess non-local adjustments. It also shows how to use non-local information from an A/B test to inform this same question, which is important for our empirical exercises.

Figuring out how to optimally locally adjust w^A of requires knowledge of $f_a\left(x|\, a^A\right)$ and $Da\left(w^A, t\right)$, which as we showed can be acquired with a local A/B test. To figure out how to best non-locally adjust w^A requires knowing $f\left(x|\, a\right)$ for all a and $a\left(w\right)$ for all w. Under the following conditions, this information can also be obtained with a single local A/B test.

Condition 1. *The score function is independent of* a: $f\left(x|\, a\right) = h\left(x\right) + ag\left(x\right)$ for some $h\left(x\right)$ and $g\left(x\right)$ satisfying $\int h\left(x\right)dx = 1$ and $\int g\left(x\right)dx = 0$.

Condition 2. *The agent has isoelastic effort costs:* $c\left(a\right) = \frac{\varepsilon}{1+\varepsilon}e^{-\beta/\varepsilon}a^{1+1/\varepsilon}$ *for some parameters* $\beta, \varepsilon \geq 0$.

The first condition ensures that knowledge of $f_a\left(\cdot|\, a\right)$ at a single effort level can be used to predict the output distribution corresponding to any effort level. It also implies that the agent's marginal incentives, $I\left(w, a\right) = \int v\left(w\left(x\right)\right)g\left(x\right)dx$, are independent of a. When this condition holds, we will drop the dependence of I on a in our notation.

The second condition implies that for any contract w, the agent's effort choice satisfies

$$\log a\left(w\right) = \beta + \varepsilon \log I\left(w\right). \tag{2}$$

A local A/B test provides the information required to estimate β and ε. It provides information on $I^A = I\left(w^A, a^A\right)$ as well as $Da\left(w^A, w^B\right)/DI\left(w^A, w^B\right)$, and the agent's elasticity of effort with respect to marginal incentives is

$$\varepsilon = \frac{Da\left(w^A, w^B\right)}{DI\left(w^A, w^B\right)}\frac{I^A}{a^A}.$$

The coefficient β can be constructed using this information as well: $\beta = \log a^A - \varepsilon \log I^A$.

In order to state the principal's problem of how best to adjust a status quo contract, we will introduce some notation. Given a contract w and a function $q\left(w\right)$, define the **increment**

of q in the direction t by $\Delta q(w,t) = q(w+t) - q(w)$. We will now revisit the agent's problem, under the assumption that Condition X1 is satisfied. Given contract w, he solves

$$u(w) = \int v(w(x))\, h(x)\, dx + \max_a \{aI(w) - c(a)\}.$$

Define the agent's optimal effort given marginal incentives I by $\tilde{a}(I)$. The following lemma parallels Lemma 1 and shows how the agent's utility and effort change in response to a non-local adjustment to w in the direction t.

Lemma 4. *Suppose Condition 1 is satisfied. Adjusting a contract w in the direction t changes the agent's utility by*

$$\Delta u(w,t) = \int \left[v(w(x) + t(x)) - v(w(x))\right] h(x)\, dx + \int_{I(w)}^{I(w+t)} \tilde{a}(I)\, dI$$

and changes his effort by

$$\Delta a(w,t) = \int_{I(w)}^{I(w+t)} \frac{d\tilde{a}(I)}{dI} dI,$$

where $d\tilde{a}(I)/dI = c''(\tilde{a}(I))$.

Lemma 4 characterizes the relevant aspects of the agent's problem. The first part of the lemma follows from the integral form of the envelope theorem. The second part of the lemma shows that the agent's behavioral response to adjusting the contract depends on the entire schedule of how his effort changes as his marginal incentives change. Under Condition X1, the principal therefore needs information on the function $d\tilde{a}/dI$ for each contract w. Condition X2 ensures that this function is entirely pinned down by knowledge of a single point as well as its slope. It would also be entirely pinned down by knowledge of any two points: it can therefore be identified from *any* A/B test, which will be relevant in our empirical exercise in Section 5. This property holds generically for any two-parameter family of effort-cost functions. [insert footnote on linear extrapolations]

Let us now define how the principal's profit changes as she adjusts w in direction t.

$$\Delta\pi(w,t) = \left(m - \int w(x)\, g(x)\, dx\right) \Delta a(w,t) - \int t(x) \left(h(x) + (a(w) + \Delta a(w,t))\, g(x)\right) dx.$$

The principal's problem of choosing an optimal non-local adjustment t to the status quo contract w^A is therefore

$$\max_t \Delta\pi\left(w^A, t\right) \text{ subject to } \Delta u\left(w^A, t\right) \geq 0. \tag{Adj}$$

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)

In practice, the program (Adj) is solved using a two-step procedure reminiscent of Grossman and Hart (1983). In the first step, we fix Δa and solve for the cost-minimizing adjustment t that satisfies $\Delta a = \Delta a\left(w^A, t\right)$ and $\Delta u\left(w^A, t\right) \geq 0$. In the second step, we choose the optimal Δa. The first stage problem can be transformed into a convex program by transforming the principal's choice from the function t to the function $V = v\left(w^A + t\right)$.

5 Empirical Validation

The goal of this section is to test the predictions of our model and to illustrate how the techniques developed in the previous section can be applied. 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.[5]

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

[5]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 the contract affects effort, and consequently the principal's profit. Exploiting the fact that this dataset contains several different treatments, in Section 5.1, we pick any pair of treatments, assume they constitute 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 assumption about the subjects' (common) marginal utility function.

In Section 5.2, again taking any pair of treatments to constitute the principal's A/B test, we solve for 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 the 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 various dimensions, such as their ability to perform in this task or in their willingness to do so, 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 (RRA) 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 4, for any contract w, the corresponding marginal benefit of effort $M(w) = \int w^{1-\rho} \widehat{f}_a dx \,/\, (1-\rho)$.[6] Then it follows from Assumption 4 and (2) that

$$a(w) = \left[\frac{\int w^{1-\rho} \widehat{f}_a dx}{\int w^{1-\rho} \widehat{f}_a dx} \right]^{\epsilon} a(w_A) \text{ , where } \epsilon = \frac{\log\left(a(w_A)/a(w_B)\right)}{\log\left(\int w_A^{1-\rho} \widehat{f}_a dx / \int w_B^{1-\rho} \widehat{f}_a dx\right)} \tag{3}$$

and $\beta = \log\left(a(w_A)\right) - \epsilon \log\left(\int w_A^{1-\rho} \widehat{f}_a dx/(1-\rho)\right)$.[7] In addition, to the "logarithmic model" described above, we also present the predictions from the "linear model" described in footnote 7. In this case, for any contract w, the model predicts

$$a(w) = a(w_A) + [a(w_B) - a(w_A)] \frac{\int (w^{1-\rho} - w_A^{1-\rho}) \widehat{f}_a dx}{\int (w_B^{1-\rho} - w_A^{1-\rho}) \widehat{f}_a dx} \,. \tag{4}$$

We denote the effort prediction corresponding to w_i using (3) and (4) by $\widehat{a}_{log}(w_i)$ and $\widehat{a}_{lin}(w_i)$, respectively. (For expositional convenience, we suppress the dependence of $\widehat{a}_{log}$ and $\widehat{a}_{lin}$ 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

[6] 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.

[7] 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 (3) 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.

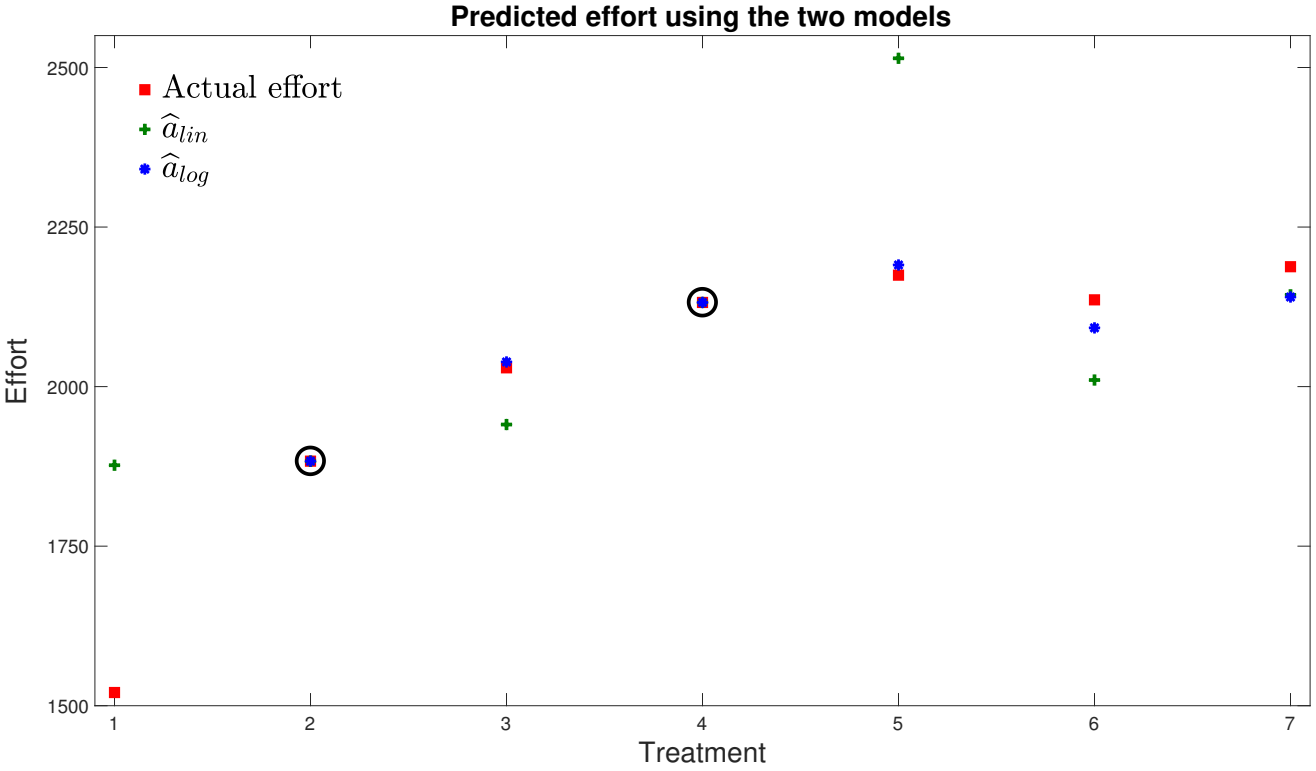


Figure 1: Predicted versus actual effort assuming that the principal has data for treatments 2 and 4, and the coefficient of RRA $\rho = 0.3$.

in footnote 7.) 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 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 worst-case APE as

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

respectively, where $k \in \{lin, log\}$. For example, given an A/B test (the notation of which is suppressed), Mean APE$(\widehat{a}_{log})$ measures the average APE when we use the logarithmic model to predict the effort corresponding to every contract other than the two comprising the A/B test, and w_1. (For the reason described above, we ignore A/B tests that involve w_1.) Figure 2 presents the mean and the worst-case APE for each of the 15 available A/B tests. For example, treatment pair 3-6 refers to the case in which the principal's A/B test comprises

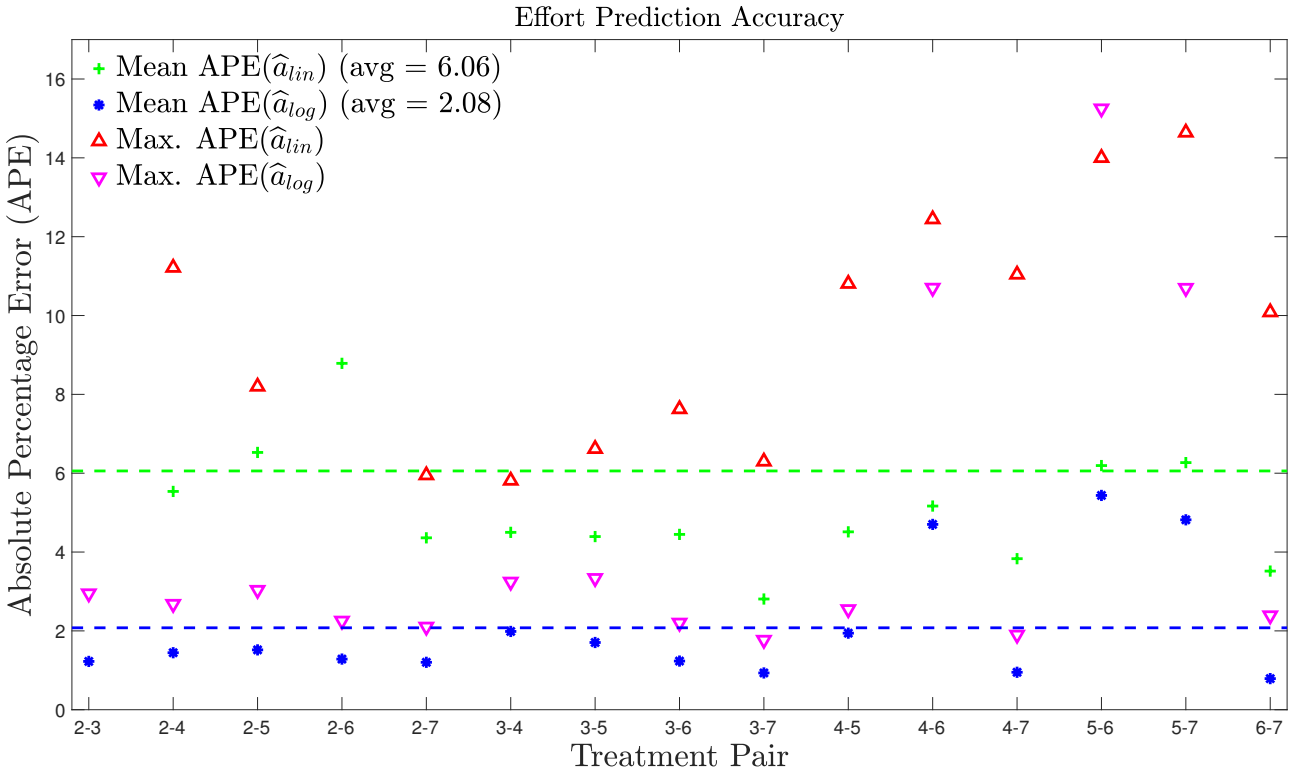


Figure 2: Effort prediction accuracy (coefficient of RRA $\rho = 0.3$).

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 elasticity from (3), $\epsilon < 0.05$ 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 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.

We make two remarks. The first is to remind the reader that $\{w_1, .., w_5\}$ are linear contracts, while $\{w_6, w_7\}$ are bonus contracts, and note that the model generates remarkably precise predictions across these two different contract classes. Second, we discuss why the A/B tests comprising $\{w_4, w_6\}$, $\{w_5, w_6\}$, and $\{w_5, w_7\}$ generate markedly poorer predictions compared to the other tests. Let us take $\{w_4, w_6\}$, and 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 (**??**), 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 (**??**), 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, so the numerator of (**??**) 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 (Adj), we obtain the following prediction for the principal's profit if the status quo contract, w_A, is replaced by w:

$$\widehat{\pi}_k(w) = m\widehat{a}_k(w) - \int w\left(\widehat{f} + [\widehat{a}_k(w_j) - a(w_A)]\,\widehat{f}_a\right) dx\,, \tag{5}$$

where $k \in \{lin, log\}$. Figure 3 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 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.

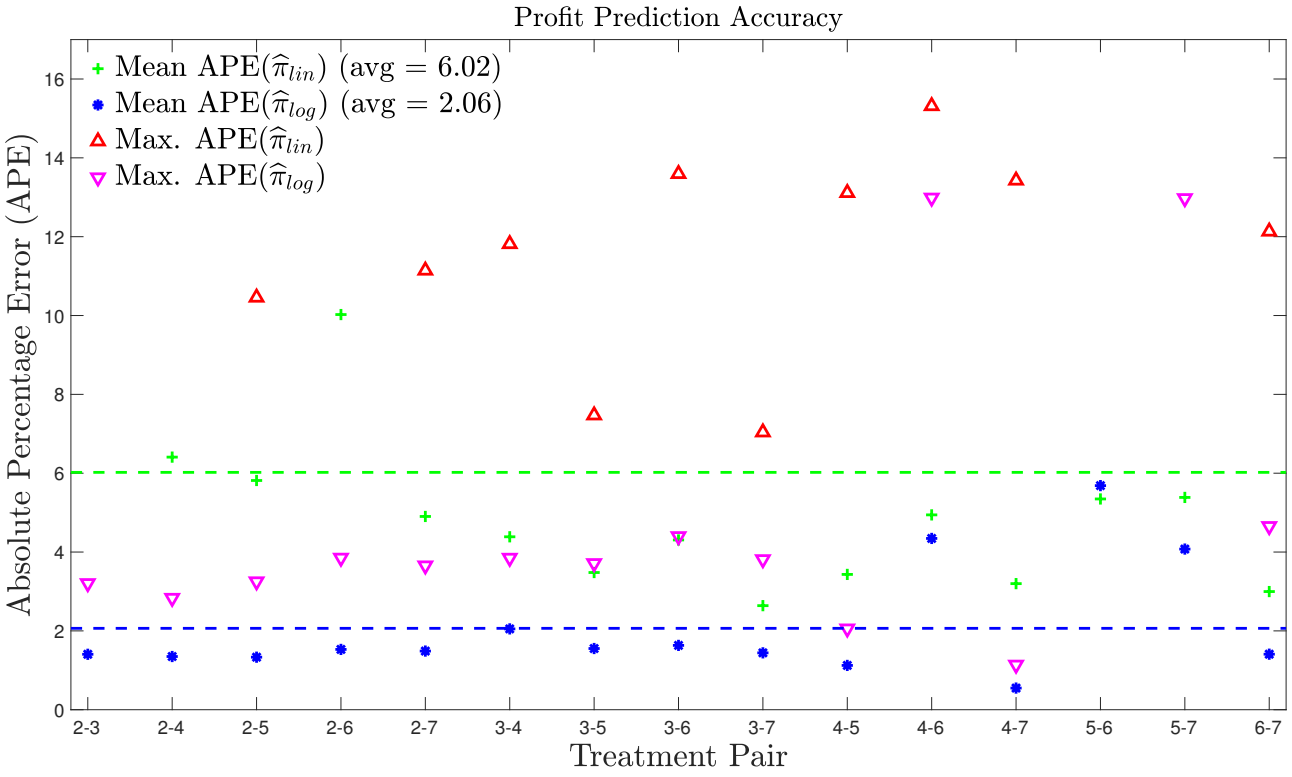


Figure 3: Profit prediction accuracy ($m = 0.2$ and coefficient of RRA $\rho = 0.3$).

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 (Adj) for all Δa, and then solving (**??**) 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).[8] Then, we define for every x,

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

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

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

Now we define $f(\cdot|a_1) \equiv f^1(\cdot)$, and recursively, for all $a \in (a_1, 2200]$, $f(\cdot|a) \equiv f(\cdot|a_1) + \int_{a_1}^{a} f_a(\cdot|s)ds$.[9] Figure 4 illustrates the empirical cumulative distribution function and $f(\cdot|a)$ for $a \in \{a_2, a_3, a_4, a_6\}$ in the left and the right panel, respectively.

Towards estimating the 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 chooses a such that

$$\int v(w(x)) f_a(x|a)dx + I = c'(a)\,, \tag{6}$$

where I is a parameter that captures indirect incentives or intrinsic motivation.[10] Table 2 reports the estimates for unknown parameters using nonlinear least squares minimization. Moreover, we assume that the principal's marginal profit, $m = 0.2$.

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

[9] 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 $a \in [a_1, 2200]$.

[10] See Appendix A.2 for a formal treatment of this case.

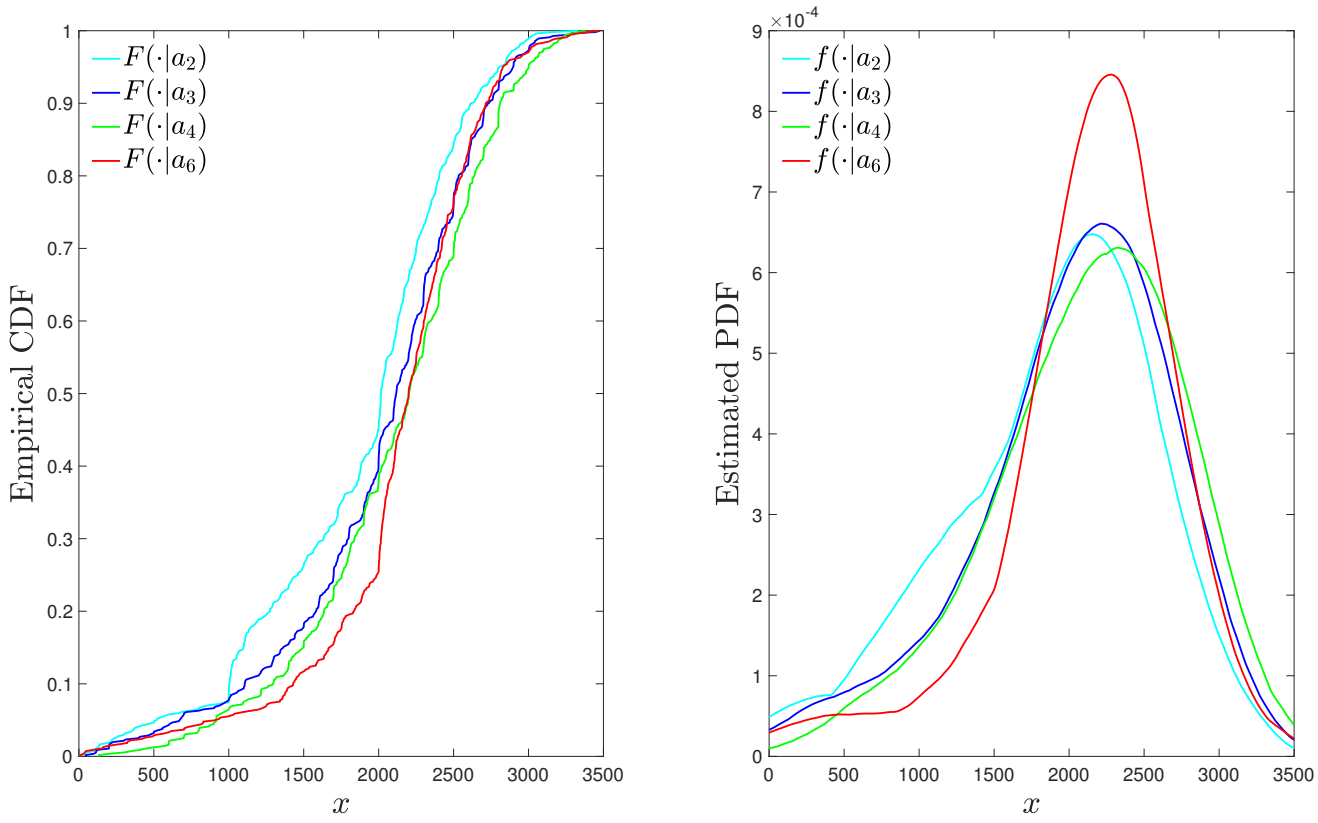


Figure 4: Empirical CDF and estimated pdf (using the triweight kernel) corresponding to treatments 2, 3, 4 and 6

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

Step 2(a): Optimal Contract (Benchmark)

We now pick an arbitrary pair of treatments (excluding treatment 1 for the reasons explained in footnote 7) 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 \$1, 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.[11]

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.)

Given these parameters, we solve (Adj) subject to (**??**), (**??**), and the two additional constraints (*i.e.*, that $w(x) \geq 100$ and increasing for all x) 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 $\widehat{w}^*$.[12] To evaluate its profit, denoted $\widehat{\Pi}^*$, we use the estimated model from Step 1.

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 5 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, evaluated using the model estimated in step 1. The contract $\widehat{w}^i_{costmin}$ for $i = A$ ($i = B$) is the solution of (Adj) when $\Delta a = 0$ ($\Delta a = a(w_B) - a(w_A)$). This corresponds to the cost-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.

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

[12] 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).

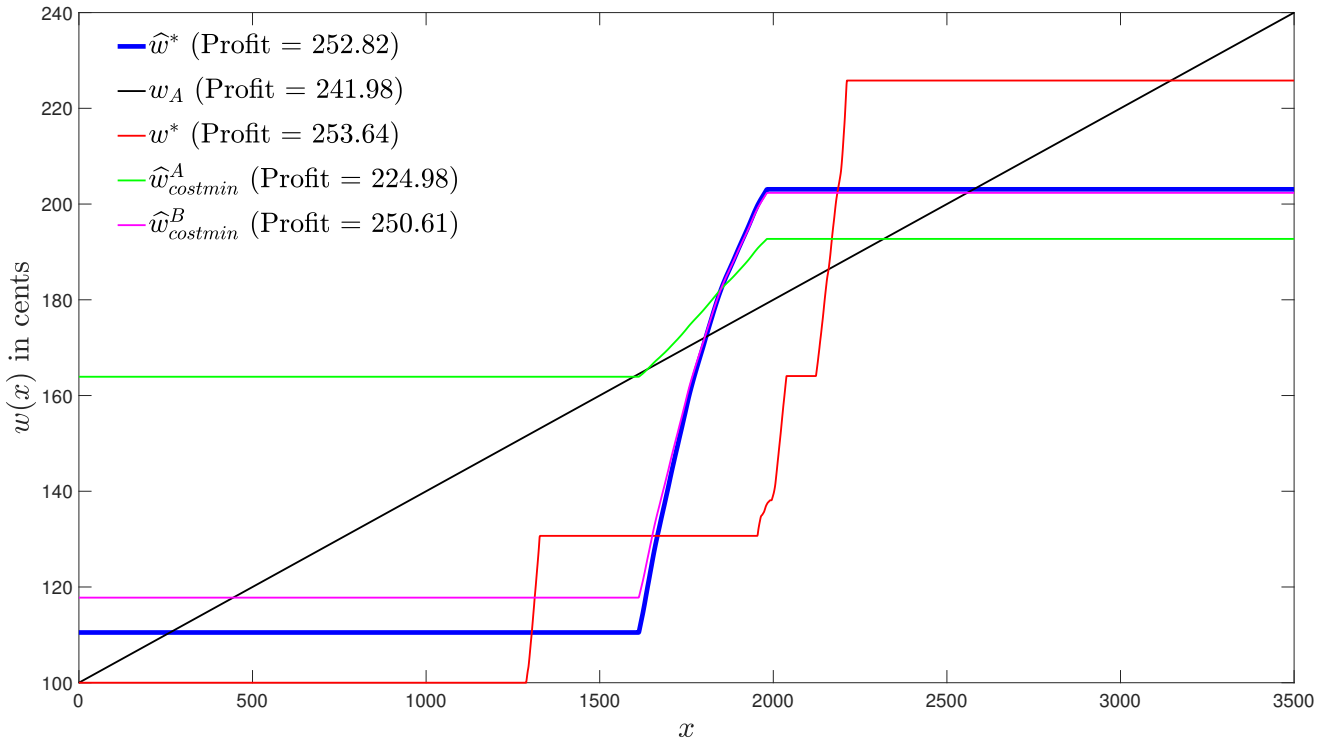


Figure 5: Optimally perturbed contract and benchmarks for the case in which the A/B test comprises treatments 4 and 7 ($m = 0.2$ and coefficient of RRA $\rho = 0.3$)

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 6 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 of 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 of the optimally perturbed contract that solves (Adj) using blue pentagrams.[13]

Averaging across all A/B tests, the optimally perturbed contract that solves (Adj) achieves 97.04% of the profits of the corresponding optimal contract. Meanwhile, the status quo contract and the best cost-minimizing contract achieves 94.72% and 95.24% of the profits of

[13] 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. See our discussion at the end of Section 5.1.

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 (Adj), $\widehat{w}^*$, 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.

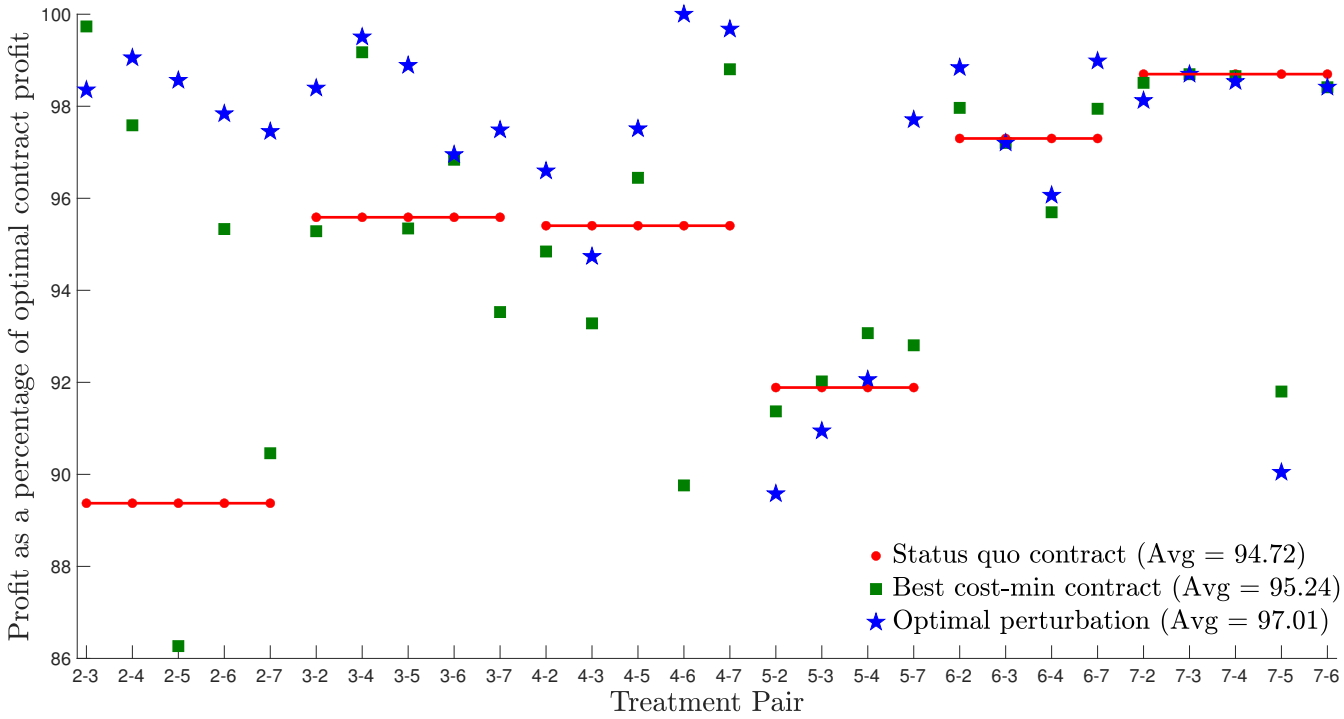


Figure 6: Profits of optimally perturbed contract relative to optimal contract and other benchmarks for every A/B test ($m = 0.2$ and coefficient of RRA $\rho = 0.3$)

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 three extensions of the algorithm presented in Section **??**. First, we consider the case in which the firm offers a contract to a group of heterogeneous workers. Second, we consider the possibility that the worker is also motivated by factors beyond direct financial incentives, such as the prospect of a promotion, prestige, the threat of firing, or is intrinsically motivated. Finally, we discuss how our algorithm can be adapted if the firm restricts attention to a particular parametric class of contracts.

A.1 Heterogeneous Abilities

So far, we have assumed that the firm 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 workers with heterogeneous effort costs.

Let $\theta \in \mathbb{N}$ denote the type of each worker. We assume that workers with different types have different costs of effort but are otherwise identical. If an type-θ worker, 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 Conditions 1-2, thus assuming that $f_a(\cdot|a)$ does not depend on a, and the cost of a type-θ worker 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 worker when offered contract w. We assume that the firm 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 workers 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 firm can estimate the constants c_0^θ and ϵ^θ by solving the following system of linear equations:

$$\log a^\theta(w_i) = \beta^\theta + \epsilon^\theta \log M(w_i) \quad \text{for } i \in \{A, B\}\,. \tag{7}$$

We will argue that under the assumptions described above, the firm'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 (Adj), (**??**), and (**??**). Let us explain each expression. First, for any given target effort level Δa^1, ($\widehat{IC}^1$) follows from (7) 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 (**??**) and (Adj), after incorporating the appropriate θ-superscripts.

This program can be solved using the same two-stage approach as (Adj): First, for every Δa^1, one finds the contract that motivates each type-θ worker 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 firm 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 firm 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 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 **??**, we maintain Assumptions 4-4, and additionally assume that I does not depend on effort. Then, for any contract w, the model predicts that the following relationship is satisfied:

$$\log a(w) = \beta + \epsilon \log\left[M(w) + I\right], \tag{8}$$

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

Turning to the firm's problem, it is straightforward to verify that (Adj) and (**??**) are unchanged when we incorporate indirect incentives. Using (8), it follows that for given I, if the principal wants to induce some effort $\widehat{a} + \Delta a$, then the contract must satisfy

$$\int v(w)\widehat{f}_a dx + I = \left(\frac{\widehat{a} + \Delta a}{\widehat{a}}\right)^{1/\epsilon}\left[\int v(w_A)\widehat{f}_a dx + I\right]. \tag{9}$$

To find the optimally revised contract, given our assumptions, the firm solves (Adj) subject to (**??**) and (9).

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.[14]

In this section, we discuss how the methodology in Section **??** can be applied when the firm 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 $w_\alpha(x) = \alpha_1 + \alpha_2 \mathbb{I}_{\{x \geq \alpha_3\}}$.

[14]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).

It is straightforward to verify that the firm's problem still solves (Adj) subject to (**??**) and (**??**), 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.

B Robustness Tests

In this section, we present a number of exercises to evaluate the robustness of our findings.

B.1 Prediction Accuracy vs. Assumptions

An important tenet of our framework is that the firm *knows* the worker's marginal utility function. In the empirical exercises in Section 5, we assumed that each worker's utility exhibits constant RRA with coefficient $\rho = 0.3$. To evaluate how the prediction accuracy of the logarithmic model (described in Section **??**) depends on this assumption, we vary the coefficient of RRA, ρ, from zero to one. We also consider the assumption that subjects have quadratic utility, and so $v'(\omega) = A - 2B\omega$, where we normalize $A = 10^3$ and vary $B \in [10^{-3}, 1]$.[15] Figure 7 illustrates the Mean APE averaged across the 15 treatment pairs as a function of ρ. Notice that the prediction accuracy is essentially invariant to whether the worker's utility function is assumed to exhibit constant RRA or is quadratic, as well as to the assumed coefficient of risk aversion.

In Section 5, we ignored A/B tests that include w_1, the treatment which pays a flat participation fee. As discussed in footnote 7, the worker's marginal incentives under this contract, $M(w_1) = 0$, and so the logarithmic model presented in (3) is not defined in that case. Considering that most A/B tests of incentive pay in the literature comprise a treatment with no pay for performance (see, for example, Lazear (2000), Shearer (2004), and others), this is potentially troublesome. Together with the fact that workers exert positive effort even if they receive no explicit pay for performance, this tension can be resolved by incorporating an additional parameter, I, which captures the motivation due to factors other than direct monetary incentives, such as the prospect of a promotion, the threat of firing, et cetera. Let us assume that the marginal benefit of effort is additive in direct and indirect incentives, and so faced with a contract w, a worker chooses his effort such that $c'(a) = M(w)+I$, as described in Appendix A.2. To determine an optimal perturbation, the firm must recover three unknown parameters from (8): β, ϵ, and I. To do so, it must either have data corresponding to at

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

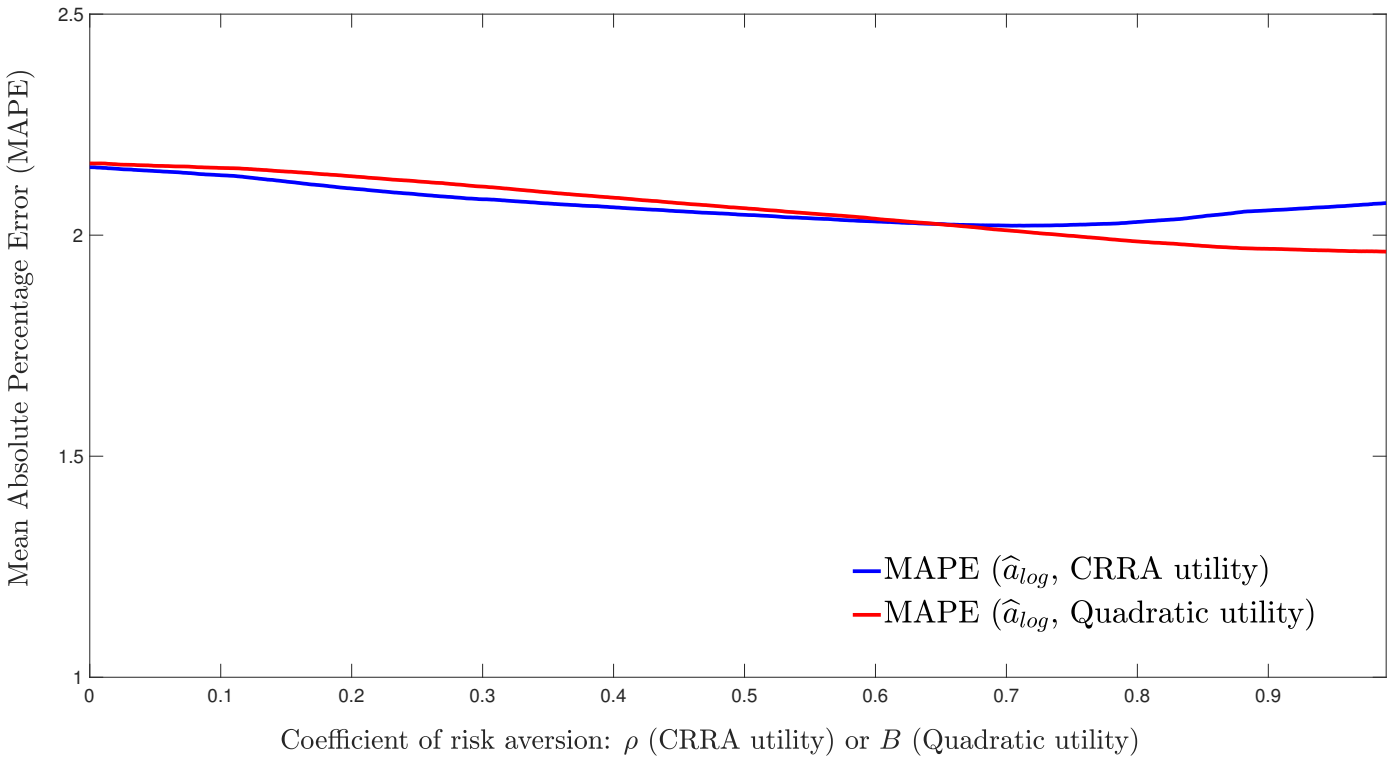


Figure 7: Effort prediction accuracy as a function of the firm's assumption about the worker's marginal utility function v'.

least three contracts, or it might assume an arbitrary value for I, then estimate β and ϵ, and ultimately predict effort using (8). Figure 8 illustrates the prediction accuracy of our model averaged across all A/B tests that include w_1 as a function of the assumed value for I when the firm's A/B test comprises w_1 and w_i for some $i \in \{2, .., 7\}$. Observe that the mean and the worst-case APE is below 5% and 12%, respectively, as long as $I \lesssim 10^{-5}$, but then rises rapidly for larger values of I.

B.2 Performance of Perturbed Contract vs. Assumptions

Figure 9 illustrates how the performance of the algorithm given in (Adj) varies with the principal's profit margin, m. Across the values of m evaluated, the optimal perturbation increases profits relative to the status quo contract by approximately 2%, and closes around half of the profit-gap to the optimal contract.

In Figure 6, we considered the case in which the worker's utility function exhibits constant relative risk aversion with coefficient $\rho = 0.3$, and the firm correctly assumes so. Figure 10 illustrates how much of the profit gap between the status quo contract and the opti-

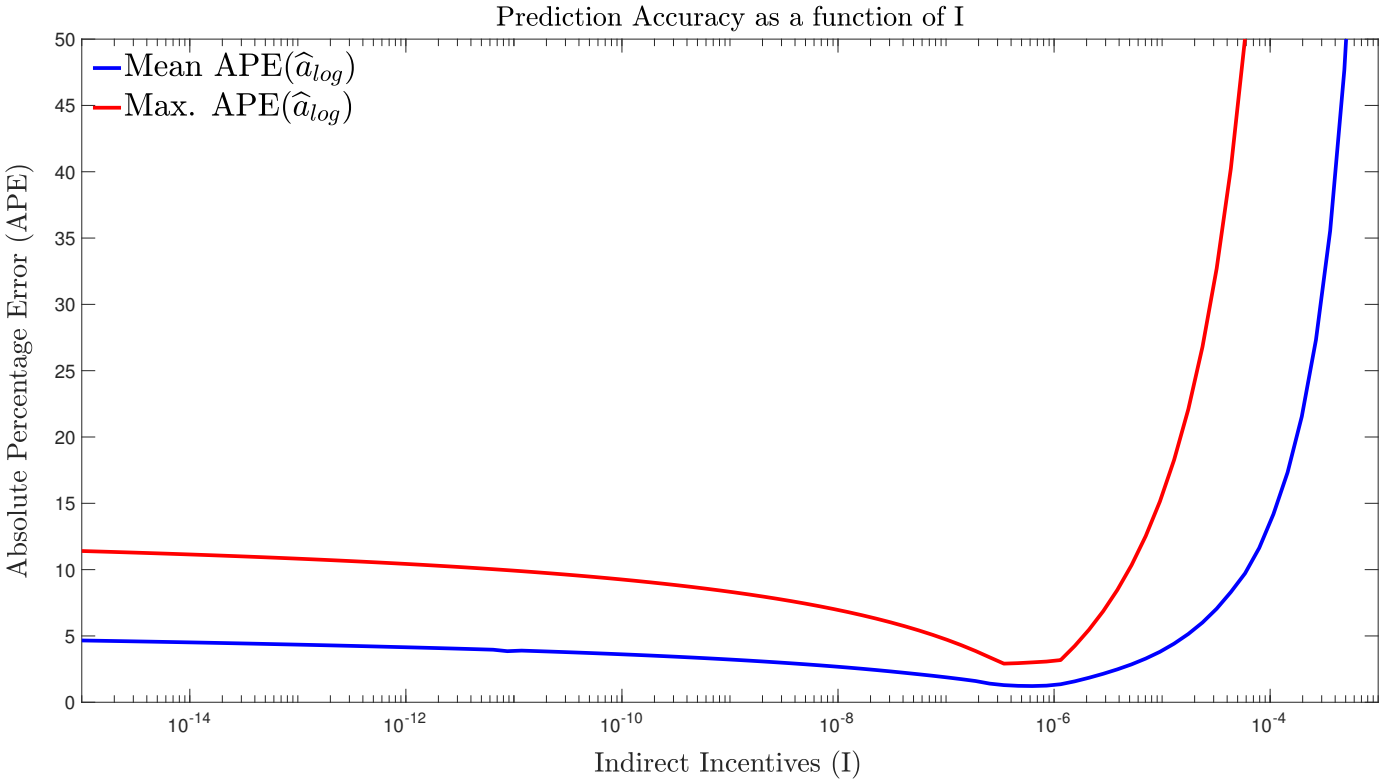


Figure 8: Effort prediction accuracy for A/B tests featuring w_1 as a function of the principal's assumption about the magnitude of indirect incentives, I.

mal contract, the optimally perturbed contract captures, averaged out across all treatment pairs, assuming that the firm makes an incorrect assumption about the worker's coefficient of (constant) RRA. Observe that the performance of the optimal perturbation is relatively invariant to the firm's assumption about ρ. Moreover, in all cases except for the one in which $\rho = 0.1$, the optimal perturbation captures about half of the profit gap between the status quo contract and the optimal contract.

In Figure 11, we vary the coefficient of RRA used to estimate the parameters of the model, meanwhile supposing that the firm assumes that $\rho = 0.3$ in all cases. Observe that the performance of the optimal perturbation is essentially invariant to the coefficient of RRA. Indeed, the estimated parameters are little affected as we vary the coefficient of RRA.

C Proofs

Lemma 2. *Take any two production environments* $P = (f, c)$ *and* $\tilde{P} = \left(\tilde{f}, \tilde{c}\right)$ *satisfying* $f^A = \tilde{f}^A$ *and* $f_a^A = \tilde{f}_a^A$ *and* $Da\left(w^A, t \middle| P\right) = Da\left(w^A, t \middle| \tilde{P}\right)$ *for all* t. *Then* t^* *solves*

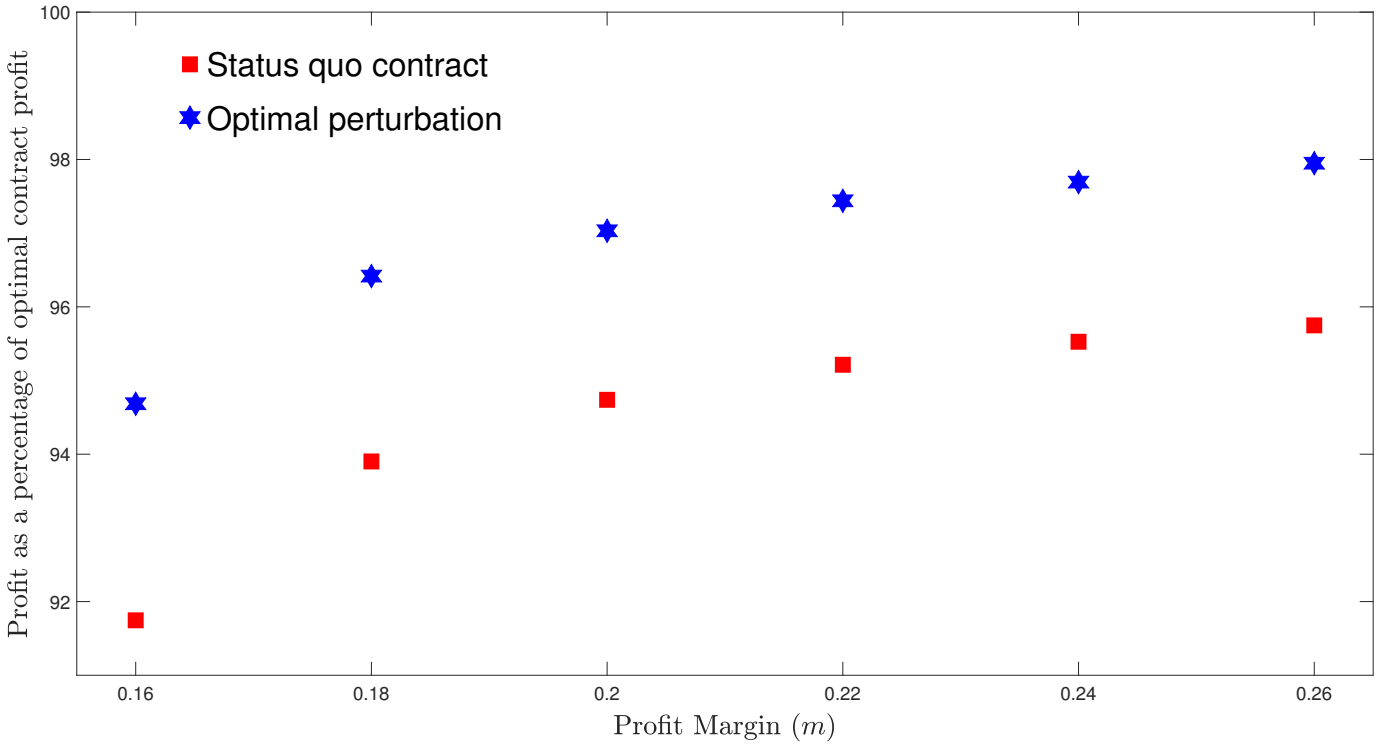


Figure 9: Profits of the optimally perturbed contract relative to the optimal contract averaged across every A/B test as a function of the firm's profit margin, m.

$(Adj_{local} - P)$ *if and only if it solves* $\left(Adj_{local} - \tilde{P}\right)$.

Proof of Lemma 2.
Immediate. □

Proof of Lemma 3.
Note that

$$\begin{aligned} LAB\left(w^A, w^B \middle| P\right) &= \left(f^A, f_a^A, Da\left(w^A, w^B \middle| P\right)\right) \\ LAB\left(w^A, w^B \middle| \tilde{P}\right) &= \left(\tilde{f}^A, \tilde{f}_a^A, Da\left(w^A, w^B \middle| \tilde{P}\right)\right). \end{aligned}$$

Suppose $LAB\left(w^A, w^B \middle| P\right) = LAB\left(w^A, w^B \middle| \tilde{P}\right)$. It is immediate that $f^A = \tilde{f}^A$ and $f_a^A = \tilde{f}_a^A$. Next, note that for all t,

$$DI\left(w^A, t \middle| P\right) = \int tv'\left(w^A\right) f_a = \int tv'\left(w^A\right) \tilde{f}_a = DI\left(w^A, t \middle| \tilde{P}\right).$$

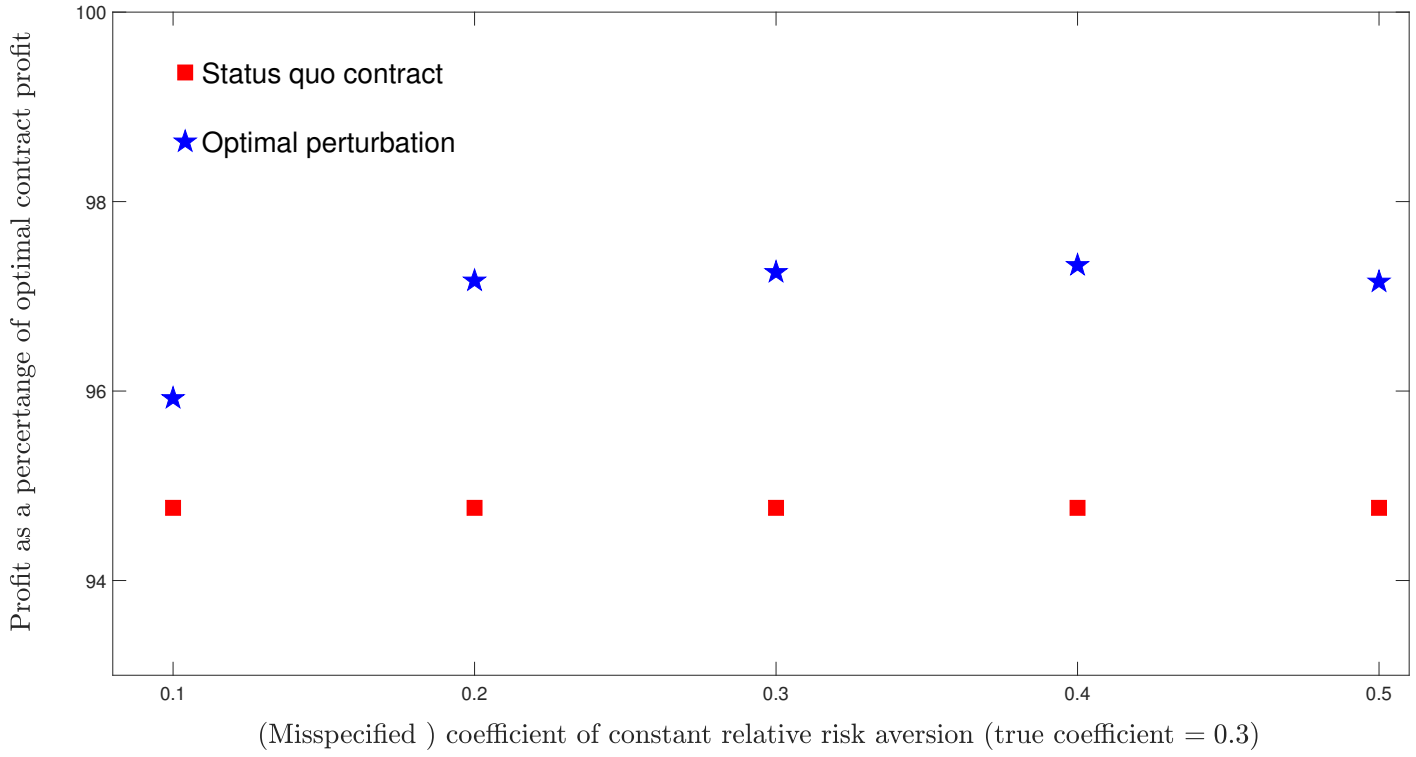


Figure 10: Profits of the optimally perturbed contract relative to the optimal contract averaged across every A/B test as a function of the firm's misspecified coefficient of relative risk aversion when $m = 0.2$ and the true coefficient of RRA equals 0.3.

By Lemma 1,

$$\begin{aligned} Da\left(w^A, t \middle| P\right) &= \frac{Da\left(w^A, w^B \middle| P\right)}{DI\left(w^A, w^B \middle| P\right)} DI\left(w^A, t \middle| P\right) \\ Da\left(w^A, t \middle| \tilde{P}\right) &= \frac{Da\left(w^A, w^B \middle| \tilde{P}\right)}{DI\left(w^A, w^B \middle| \tilde{P}\right)} DI\left(w^A, t \middle| \tilde{P}\right), \end{aligned}$$

so $Da\left(w^A, t \middle| P\right) = Da\left(w^A, t \middle| \tilde{P}\right)$ for all t if and only if $Da\left(w^A, w^B \middle| P\right) = Da\left(w^A, w^B \middle| \tilde{P}\right)$. □

Proof of Proposition 1. Let $\lambda \geq 0$ and $\nu \geq 0$ denote the dual multipliers associated with the

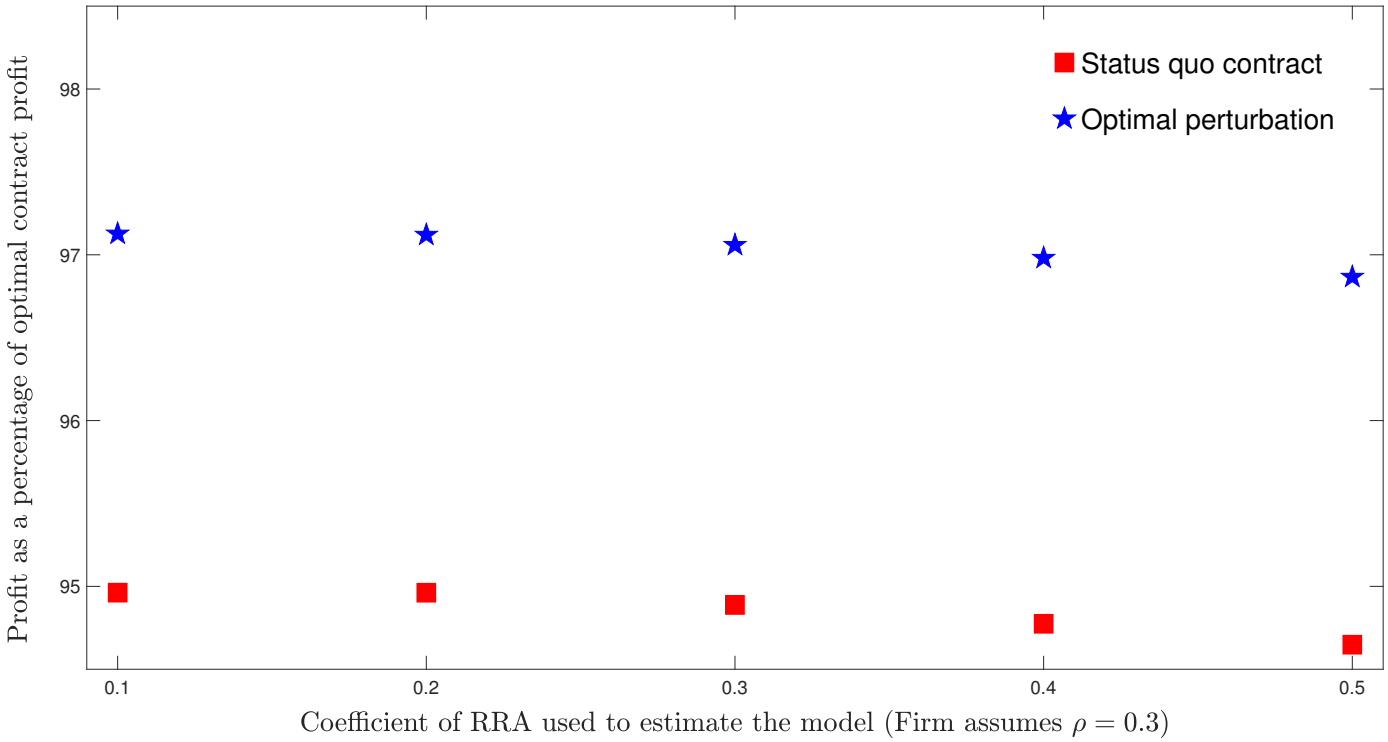


Figure 11: Profits of the optimally perturbed contract relative to the optimal contract averaged across every A/B test as a function of the coefficient of RRA used to estimate the parameters of the model, meanwhile supposing that the firm assumes that $\rho = 0.3$.

first and second constraint in (**??**), respectively. The Lagrangian

$$L(\lambda, \nu) = \max_t \left\{ \nu + \int \left[t \left(\lambda v'(w_A) \widehat{f} + \mu v'(w_A) \widehat{f}_a - \widehat{f} \right) - \nu t^2 \right] dx \right\}. \tag{10}$$

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.[16]

Next, we pin down the optimal multipliers λ and ν, by turning to the dual problem, and

[16] If $\nu = 0$, then the integrand of (10) 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 $L(\lambda, 0) = 0$.

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

$$\lambda^* = \max\left\{0, \frac{\int \left(\widehat{f} - \mu v'(w_A)\widehat{f}_a\right) v'(w_A)\widehat{f}dx}{\int \left(v'(w_A)\widehat{f}\right)^2 dx}\right\} \tag{11}$$

and

$$\nu^* = \frac{1}{2}\sqrt{\int \left[\left(\lambda^*\widehat{f} + \mu\widehat{f}_a\right) v'(w_A) - \widehat{f}\right]^2 dx}\,. \tag{12}$$

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 (**??**). 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 (11) and (12) 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 (**??**). Weak duality implies that $L(\lambda^*, \nu^*) \geq \Pi^*$. Moreover, it is straightforward to verify that $t\,(\lambda^*, \nu^*)$ is feasible for (**??**), and λ^* and ν^* is strictly positive if and only if the respective (primal) constraint binds. This implies that the objective of (**??**) 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 (**??**).

To complete the proof, we show that w_A is locally optimal if and only if (**??**) 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 (10) 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. □