Abstract

This paper aims to improve the practical applicability of the classic theory of incentive contracts under moral hazard. We establish conditions such that the information provided by an A/B test of incentive contracts is a sufficient statistic for the question of how best to 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.
1 Introduction

Firms and organizations throughout the economy now understand that there is a lot to learn from experimentation—they regularly use it to inform product design, pricing, advertising, and many other facets of their product-market strategies. Equally critical to the survival of any organization, however, is the management of compensation and reward structures: How should people be rewarded for outcomes? This can be a challenging question to answer—even in theory—and it has largely evaded recent trends in data-driven decision-making. This paper shows that under some mild and reasonable assumptions about the way people respond to incentives and value monetary rewards, simple experimentation coupled with a few basic theoretical insights can lead us a long way towards answering it.

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 piece rate for doing so. You have access to 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 $\alpha$, and your worker’s average sales are $a$. Your expected profits are therefore $\Pi = (m - \alpha)a$. If you were to marginally increase your piece rate, the effect on your profits would be

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

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. You do not, however, know your workers’ behavioral response, $da/d\alpha$, to an increase in the piece rate. Given observational data alone, figuring out this behavioral response requires knowing a lot about the problem your workers face: How much do they value money? What are their effort costs? And 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; see, for example, Chetty (2009), and the references therein.

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 adjustments. 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 \( f(\cdot|a) \), 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 adjustment \( dw \) to the contract. This adjustment directly affects the expected wage bill by \( E[dw] \) and leads the agent to change his effort level by some amount, \( da \). The total effect on the principal’s profits is therefore

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

which is the appropriate generalization of (1) to nonlinear contracts\footnote{We write \( f_a \) to denote the derivative of \( f(\cdot|a) \) with respect to \( a \), and we suppress the dependence on \( x \) and \( a \) for notational simplicity.}. The main challenge to figuring out the best marginal adjustment 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 adjusted. Our main lemma shows that, given knowledge of the agent’s 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 adjustment 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 cost of exerting additional effort equals his marginal incentives, which are given by $I = \text{Cov}(v(w), f_a/f)$. 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, $da/dI$, is independent of the adjustment to the contract that led to the change in marginal incentives. His behavioral response to a marginal adjustment to the contract, $\tilde{dw}$, can therefore be expressed as $\tilde{da} = (da/dI)\tilde{dI}$. Predicting how the agent will respond to an adjustment to the contract therefore requires information about how he will respond to a change in his marginal incentives, and it requires information about how the adjustment to the contract affects his marginal incentives.

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 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 adjustment to the status quo contract affects the agent’s marginal incentives, $\tilde{dI}$, and therefore the agent’s effort choice $\tilde{da} = (da/dI)\tilde{dI}$. 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 adjustment to the status quo contract and serves as a sufficient statistic for the question of how best to marginally adjust the status quo contract. This sufficient-statistic result is our main conceptual contribution. We then show that the problem of how best to locally adjust 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 convex program.

The second important issue that the above example sidestepped was the question of how to predict the effects of non-local adjustments 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 adjustment to the status quo contract. In doing so, we provide an algorithm for figuring out how to use this information to optimally
adjust 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 several 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 exercise asks the question of whether subjects' average output varies in the way our model predicts with our measure of the subjects' marginal incentives. We take the data from two treatments within the same class, that is, data from two piece-rate treatments or two bonus treatments. We 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. For each such pair, we predict the mean output in each of the remaining four treatments and compare it to the actual average output. A/B tests using piece-rate contracts predict the performance in the other piece-rate-contract treatments well: the average absolute percentage error for such predictions is 0.66%. A/B tests using piece-rate contracts also predict the performance in bonus contracts well, and vice versa: the average absolute percentage error for such predictions is 2.28%. The correlation between all of our predictions and actual performance is 0.94. Moreover, our predictions for a given treatment are similar no matter which A/B test we use to make our predictions.

Our second empirical exercise assesses the performance of the contract generated by our algorithm. We use data from five 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 treatments within the same class, supposing that one is the status quo contract and the other is the test contract, and we use our algorithm to construct the optimally adjusted contract. We will define the realized gains of an adjustment to be the difference between the profit corresponding to the adjusted contract and the profit of the status quo contract. We will define the maximum gains available to be the difference between the profits of the optimal and the status quo contract. Averaging across all A/B tests, the realized gains are equal to approximately 68% of the maximum gains. About two-fifths of the gap between realized and maximum gains is due to the principal implementing a too low effort level, while the remainder is due to the optimal adjustment implementing an effort
level at too high a 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 agent’s effort is multidimensional, to settings where the firm employs heterogeneous agents, and to settings where the principal is constrained to choosing from a parametric class of contracts, such as linear or bonus 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 non-contractible action. 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.

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. Others study work on more complex tasks that are amenable to the multitasking problem; see, for example, Holmström and Milgrom (1991). Gibbs et al. (2017) exploit 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

\[^2\] Exceptions include 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 parameter of the model.\[^3\] Oettinger (2001) and Fehr and Goette (2007) find a positive effect of commissions on sales for stadium vendors and on productivity for bicycle messengers in Zürich, respectively. Bandiera et al. (2007) and Bandiera et al. (2009) measure the effect of introducing performance pay for managers on their subordinates’ productivity. Guiteras and Jack (2018) study 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.
to submit fewer but higher-quality ideas. 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 instruments affect certain measures of performance. The goal of this study is to bridge these literatures by exploring how an organization can combine insights from the theoretical agency literature with estimates such as those described above to improve its incentive systems.

Conceptually, this paper is related to the literature which characterizes optimal mechanisms in terms of the relevant elasticities using a variational approach. 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 monopolist-optimal quantity-discount price-menu in terms of substitution elasticities. Saez (2001) and a growing literature derives optimal income tax formulas using elasticities of earnings with respect to tax rates. See also Chetty (2009) and the references therein for an overview of this approach in other applications.

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

\footnote{Similarly, Balbuzanov et al. (2017) find that the introduction of incentives led journalists in Kenya to submit fewer, higher quality articles. Hong et al. (2018) estimate the impact of piece rates at a Chinese manufacturing firm on the quantity and quality of output.}
If the agent is paid $\omega$ 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 condition 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 that Pareto improves a status quo contract. The set of contracts we allow the principal to choose from 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 4, it will be the full set of contracts.

### 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 solving this problem. In Section 4, we will show how to extrapolate the local conditions we identify here to answer the more practical question of how best to adjust the contract non-locally.

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 $\mathcal{D}q(w, t) \equiv \lim_{\theta \to 0} [q(w + \theta t) - q(w)] / \theta$.

We will first show 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 condition. To this end, define the agent’s marginal incentives as $I(w,a) \equiv \int v(w(x)) f_a(x|a) \, dx$, where $f_a(x|a)$ is the derivative of $f(x|a)$ with respect to $a$. 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.** Locally adjusting a contract $w$ in the direction $t$ changes the agent’s utility by

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

and his effort by

$$D a(w,t) = \frac{D I(w,t)}{c''(a(w)) - \int v(w(x)) f_{aa}(x|a(w)) \, dx}, \quad (2)$$

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

The first part of the lemma shows that how the agent’s utility changes does not depend directly on his cost function. This result follows directly from the envelope theorem. The second part shows that the agent’s behavioral response depends on how the adjustment affects his marginal incentives, $D I(w,t)$, as well as on the local curvature of his problem. It also implies that $D a(w,t) / D I(w,t)$ 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 under 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 the profit differential

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

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 can now 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

\[
\max_{t: ||t|| \leq 1} \mathcal{D}\pi(w^A, t) \quad \text{subject to} \quad \mathcal{D}u(w^A, t) \geq 0, \quad (\text{Adj}_{\text{local}})
\]

where \( ||\cdot|| \) is the \( \ell^2 \) 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 the principal knows the production environment. We now show she only needs to know certain local aspects of the production environment. To do so, we will compare her problem across different production environments, and so it will be helpful to introduce the notation \((\text{Adj}_{\text{local}}-P)\) to refer to the principal’s problem \((\text{Adj}_{\text{local}})\) when the production environment is \( P \). Denote the agent’s effort choice, the output density function, and its derivative with respect to effort under the status quo contract by \( a^A = a(w^A) \), \( f^A = f(\cdot|a^A) \), and \( f^A_a = f_a(\cdot|a^A) \), respectively, and, in an abuse of notation, denote the agent’s effort differential under production environment \( P \) by \( \mathcal{D}a(w, t|P) \). The following lemma shows which aspects of the production environment are relevant for solving \((\text{Adj}_{\text{local}})\).

**Lemma 2.** Take any two production environments \( P = (f, c) \) and \( \tilde{P} = (\tilde{f}, \tilde{c}) \) satisfying \( f^A = \tilde{f}^A \), \( f^A_a = \tilde{f}^A_a \), and \( \mathcal{D}a(w^A, t|P) = \mathcal{D}a(w^A, t|\tilde{P}) \) for all \( t \). Then \( t^* \) solves \((\text{Adj}_{\text{local}}-P)\) if and only if it solves \((\text{Adj}_{\text{local}}-\tilde{P})\).

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 \( AB(w^A, w^B) \equiv (f^A, f^B) \), 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 \( LAB(w^A, w^B) \equiv (f^A, f^A_a, \mathcal{D}a(w^A, w^B)) \) 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 \( \mathcal{D}a(w^A, w^B) \neq 0 \). 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 proposition shows that the information provided by a local A/B test suffices for solving \((\text{Adj}_{\text{local}})\).

**Proposition 1.** Take any two production environments \( P = (f, c) \) and \( \tilde{P} = (\tilde{f}, \tilde{c}) \), a status quo contract \( w^A \), and an informative test contract \( w^B \). The following are equivalent:

(i) \( f^A = \tilde{f}^A, f^A_a = \tilde{f}^A_a \) and \( Da(w^A, t | P) = Da(w^A, t | \tilde{P}) \) for all \( t \).

(ii) \( LAB(w^A, w^B | P) = LAB(w^A, w^B | \tilde{P}) \).

The proof of Proposition 1 shows how the information from a local A/B test can be used to construct the necessary information for solving \((\text{Adj}_{\text{local}})\). In particular, the local A/B test can be used to compute how his marginal incentives change in response to adjusting the status quo contract in any direction; i.e., \( DI(w^A, t) \) for any \( t \). Then, using the insight from Lemma 1 that the agent’s behavioral response to a change in his marginal incentives is independent of the adjustment that led to that change, it follows that for any \( t \),

\[
Da(w^A, t) = \frac{Da(w^A, w^B)}{DI(w^A, w^B)} DI(w^A, t).
\]

The agent’s effort differential is therefore proportional to the change in his marginal incentives as the status quo contract is adjusted in the direction \( t \). To summarize, a local A/B test contains information about \( Da(w^A, w^B) \) and allows one to compute \( DI(w^A, t) \) for any \( t \). Therefore, one can also compute \( Da(w^A, t) \) for any \( t \), and consequently solve the principal’s problem, \((\text{Adj}_{\text{local}})\).

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

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

Proposition 2 characterizes the optimal local adjustment.

**Proposition 2.** Let \( w^A \) be the status quo contract. Suppose \( t^* \) solves \((\text{Adj}_{\text{local}})\). There exist \( \lambda^*, \mu^* \geq 0 \) 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 the optimal local adjustment is in the direction of a Holmström-Mirrlees-type contract, that is, it locally balances risk allocation with
incentive provision: It shifts payments from outputs where the agent has a low marginal utility of money to those where his marginal utility of money is higher. And it shifts payments towards outputs that change the agent’s marginal incentives in the profit-maximizing direction. The optimal way to balance these two considerations is determined by the coefficients $\lambda^*, \mu^*$, the exact expressions for which which are given in the proof of Proposition 2 in the appendix.

The second part of this proposition echoes the optimality conditions of Holmström (1979) and serves as a consistency check. When the contract is optimal, the coefficients $\lambda^*$ and $\mu^*$ coincide with those in Holmström (1979). The primary contribution of Proposition 1 is to show how $\lambda^*$ and $\mu^*$ change as we consider status quo contracts that are not locally optimal. In particular, $\mu^*$, 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, $\lambda^*$, is smaller when $\mu^*$ is higher.

4 Non-local Adjustments

The analysis in the previous section illustrates how local information suffices for characterizing optimal local adjustments. This section provides a method for extrapolating to assess nonlocal adjustments. It shows in particular how to use nonlocal information from an A/B test to inform this question, which is important in practice.

Figuring out how to optimally locally adjust $w^A$ requires knowledge of $f_a(x|a^A)$ and $D_a(w^A, t)$, 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 $f(x|a)$ for all $a$ and $a(w)$ for all $w$. This section will provide a pair of conditions under which this information can be obtained with a single A/B test.

**Condition 1.** The output distribution $f(x|a)$ is affine in $a$, that is, $f(x|a) = g(x) + ah(x)$ for some $g(x)$ and $h(x)$ satisfying $\int g(x) \, dx = 1$ and $\int h(x) \, dx = 0$.

This condition is useful in two ways. First, it ensures that knowledge of the output distribution, $f(\cdot|a)$, at two effort levels, say $a(w^A)$ and $a(w^B)$, is sufficient to estimate the output distribution corresponding to any other effort level. Second, it implies that this knowledge also suffices to compute $f_a(x|a) = h(x)$ and the agent’s marginal incentives, $I(w, a) = \int v(w(x)) \, h(x) \, dx$, which are independent of $a$. When this condition holds, we will drop dependence of $I$ on $a$ in our notation.
We will now revisit the agent’s problem, under the assumption that Condition 1 is satisfied. Given a contract $w$, he solves
\[
    u(w) = \int v(w(x))g(x)\,dx + \max_a \{aI(w) - c(a)\}.
\]

Another implication of Condition 1 is that the agent’s utility is concave in $a$, and so the first-order approach is valid. Define the agent’s optimal effort given marginal incentives $I$ by $\tilde{a}(I)$. The following lemma parallels Lemma 1 and characterizes the agent’s utility and effort when contract $w$ is replaced with contract $\tilde{w}$.

**Lemma 3.** Suppose Condition 1 is satisfied, and the contract $w$ is replaced with $\tilde{w}$. Then the agent’s utility
\[
    u(\tilde{w}) = u(w) + \int [v(\tilde{w}(x)) - v(w(x))]g(x)\,dx + \int_{I(w)}^{I(\tilde{w})} \tilde{a}(i)\,di
\]

and his effort
\[
    a(\tilde{w}) = a(w) + \int_{I(w)}^{I(\tilde{w})} \frac{d\tilde{a}(i)}{di}\,di,
\]
where $d\tilde{a}(i)/di = 1/c''(\tilde{a}(i))$.

Lemma 3 characterizes the relevant aspects of the agent’s problem and shows that, under Condition 1, the principal needs two pieces of information. She needs information on how the agent values the contractual adjustment, as well as how his effort changes in response to the contractual adjustment. The main observation of Lemma 3 is that this latter object does not depend directly on the adjustment being considered but instead depends only on how that adjustment affects the agent’s marginal incentives. The first part of the lemma follows from the integral form of the envelope theorem, and the second part of the lemma follows directly from the fundamental theorem of calculus.

The next condition ensures that an A/B test provides all the information required to assess how the agent will respond to adjusting the contract.

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

Condition 2 implies that for any contract $w$, the agent’s effort choice satisfies
\[
    \ln a(w) = \beta + \varepsilon \ln I(w).
\]
An A/B test provides the information required to estimate $\beta$ and $\varepsilon$. It provides information on $I^A = I(w^A)$ and $I^B = I(w^B)$, and the agent’s elasticity of effort with respect to marginal incentives is constant and so can be estimated with the arc elasticity implied by the A/B test:

$$\varepsilon = \frac{\ln a^A - \ln a^B}{\ln I^A - \ln I^B}.$$ 

The coefficient $\beta$ can be constructed using this information as well: $\beta = \ln a^A - \varepsilon \ln I^A$. Condition 2 ensures that the function $da/dI$ is fully identified by the knowledge of any two points. This property holds generically for any two-parameter family of effort-cost functions.

Let us now define the principal’s profit when she offers (some) contract $\tilde{w}$.

$$\pi(\tilde{w}) = ma(\tilde{w}) - \int [\tilde{w}(x)] [g(x) + a(\tilde{w})h(x)] dx$$

The principal’s problem of choosing an optimal nonlocal adjustment $t$ to the status quo contract $w^A$ is therefore

$$\max_{\tilde{w}} \pi(\tilde{w}) \text{ subject to } u(\tilde{w}) \geq u(w^A).$$

(Adj)

In practice, the program (Adj) is solved using a two-step due to Grossman and Hart (1983). In the first step, we fix a target effort level $a$, and we solve for the cost-minimizing contract that satisfies $a(\tilde{w}) = a$ and $u(\tilde{w}) \geq u(w^A)$. In the second step, we choose the optimal target effort. The first-stage problem can be transformed into a convex program by transforming the principal’s choice from the function $\tilde{w}$ to the function $V = v(\tilde{w})$. In general, the second-stage problem need not be a convex program. In practice, it is a one-dimensional problem that can be quickly solved numerically.

We conclude this section by formally stating the sufficient-statistic analogue of Proposition 1 in the context of non-local adjustments.

**Proposition 3.** Suppose Conditions 1 and 2 hold. Take any two production environments $P = (f, c)$ and $\tilde{P} = (\tilde{f}, \tilde{c})$, a status quo contract $w^A$, and a test contract $w^B$ for which $a(w^A) \neq a(w^B)$. The following are equivalent:

(i) $g = \tilde{g}$, $h = \tilde{h}$, $\varepsilon = \tilde{\varepsilon}$, and $\beta = \tilde{\beta}$.

(ii) $AB(w^A, w^B|P) = AB(w^A, w^B|\tilde{P})$.

Moreover, if these statements hold, then $w^*$ solves (Adj−P) if and only if it solves (Adj−P).

This proposition shows that an A/B test, in conjunction with Conditions 1 and 2, provides the necessary information to solve the principal’s problem given in (Adj).
5 An Empirical Exploration

We will now assess the quantitative implications of our model. To do so, we use data from DellaVigna and Pope’s (2017) real-effort experiment conducted on Amazon’s Mechanical Turk. In the experiment, subjects were tasked with repeatedly pressing the ‘a’ and ‘b’ keys in alternating order. They received one point for every a/b keystroke pair they managed to complete in a ten-minute period, and they were paid according to how many points they accumulated during that time. Each subject was randomly assigned to a single treatment and performed this task once.

In the treatments we focus on, subjects in different treatments were paid according to different incentive contracts. During the course of the treatment, subjects could see the incentive contract they were on, a count-down clock, a running tally of the number of keystroke pairs they had completed, as well as their accumulated earnings. We observe, for each subject, the treatment they were assigned and the number of points they accumulated.

Table 1 summarizes seven treatments. In each treatment, subjects received a $1 participation fee regardless of how many points they accumulated. In the first treatment, subjects were told only that “Your score will not affect your payment.” This corresponds to a contract $w^1(x) = 100$, where we denominate the payments in cents. We will refer to treatment 1 as the no-incentives treatment. In treatments 3 to 5, they were paid a constant amount for every one hundred points they accumulated, whereas in treatment 2, they were paid a constant amount for every thousand points they accumulated. In treatment 3, for example, they were told, “You will be paid an extra 1 cent for every 100 points.” This corresponds to a contract $w^3(x) = 100 + 0.01x$, where $x$ is the number of points achieved. We will refer to treatments 2 to 5 as the piece-rate treatments. In treatments 6 and 7, subjects received a payment if they achieved 2,000 or more points. In treatment 6, for example, subjects were told, “You will be paid an extra 40 cents if you score at least 2,000 points.” This corresponds to the contract $w^6(x) = 100 + 40I_{\{x\geq2000\}}$, where $I_{\{x\geq2000\}}$ is the indicator function for $x \geq 2000$. We will refer to treatments 6 and 7 as the bonus treatments.

We use these data to carry out two exercises. Our first exercise asks whether subjects’ average performance varies in the way our model predicts with our measure of the subjects’ marginal incentives. We use information from two treatments to predict the performance in out-of-sample treatments. The second exercise assesses the performance of the optimal adjustment generated by our algorithm relative to a benchmark that we construct from the data from the no-incentives and piece-rate treatments.

---

5 For consistency with our model, we treat $x$ as a continuous variable. Therefore, the implied incentive contracts for these treatments are an approximation.
Table 1: Experimental Treatments from DellaVigna and Pope (2017)

<table>
<thead>
<tr>
<th>Contract</th>
<th>Avg. #points</th>
<th>Std. Dev.</th>
<th>#Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-incentives</td>
<td>1521</td>
<td>31.23</td>
<td>540</td>
</tr>
<tr>
<td>Piece-rate</td>
<td>1883</td>
<td>28.61</td>
<td>538</td>
</tr>
<tr>
<td>$w^2(x) = 100 + 0.01x$</td>
<td>2029</td>
<td>27.47</td>
<td>558</td>
</tr>
<tr>
<td>$w^3(x) = 100 + 0.01x$</td>
<td>2029</td>
<td>27.47</td>
<td>558</td>
</tr>
<tr>
<td>$w^4(x) = 100 + 0.04x$</td>
<td>2132</td>
<td>26.42</td>
<td>566</td>
</tr>
<tr>
<td>$w^5(x) = 100 + 0.10x$</td>
<td>2175</td>
<td>24.28</td>
<td>538</td>
</tr>
<tr>
<td>Bonus</td>
<td>2136</td>
<td>24.66</td>
<td>545</td>
</tr>
<tr>
<td>$w^6(x) = 100 + 40I_{x\geq2000}$</td>
<td>2187</td>
<td>22.99</td>
<td>532</td>
</tr>
<tr>
<td>$w^7(x) = 100 + 80I_{x\geq2000}$</td>
<td>2187</td>
<td>22.99</td>
<td>532</td>
</tr>
</tbody>
</table>

Table 1: This table describes seven experimental treatments from DellaVigna and Pope (2017) that differed in the monetary incentives offered to the subjects. The second column describes the implied incentive contract, denominated in cents. The remaining columns describe, for each treatment, the average number of points accumulated, the standard deviation, and the number of subjects.

5.1 Predicting Out-of-Sample Experimental Results

Our results in Section 4 show how to use outcome data from two contracts to predict agents’ effort under an arbitrary contract. We will assess the accuracy and precision of such predictions by taking outcome data from two treatments, supposing one is the status quo contract, one is the test contract, and using our model to predict average performance in the remaining treatments.

Before we do so, we will discuss two aspects of DellaVigna and Pope’s (2017) experimental setting that differ from our model and that may adversely impact our predictions. 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, it is likely that subjects differ in various dimensions such as their ability or willingness to perform repetitive tasks. Each subject participated only once, so we are unable to estimate any subject-specific heterogeneity. As such, we treat subjects as being homogeneous, and we use our baseline model to make our predictions. Section 6.3 shows how to incorporate subject-specific information into the model.

We are implicitly assuming that at the outset of the experiment, each subject observes the contract he or she is offered and chooses “effort” $a$. Then the number of points he or she accumulates over the ten-minute period, $x$, is drawn from some probability distribution with mean $a$. We therefore interpret effort as being the average number of points accumulated in a particular treatment. Throughout, we will assume that Conditions 1 and 2 hold. That is, $f(x|a) = g(x) + ah(x)$ for some $g(x)$ and $h(x)$ satisfying $\int g(x)\,dx = 1$ and $\int h(x)\,dx = 0$,
and \( c'(a) = e^{-\beta/\varepsilon}a^{1/\varepsilon} \) for some parameters \( \varepsilon \) and \( \beta \). We will also assume that the agent has constant-relative-risk-aversion preferences over money, so that \( v'(\omega) = \omega^{-\rho} \). We will assume that \( \rho = 0.3 \) and assess the sensitivity of our predictions to this assumption.

Let us now outline the exercise, and then we will get into the specifics. We are going to use outcome data from two treatments—let us call them \( A \) and \( B \)—to predict average output in the remaining treatments. To do so, we will use the data from these two treatments to construct an estimate of the function \( f_a(\cdot|a) \) and the two parameters of the agent’s cost function. We will then look at a third treatment, \( C \), and predict the agent’s marginal incentives under that treatment. This exercise will give us a prediction for average output in treatment \( C \). We will then compare these predictions to the actual average output in that treatment.

Specifically, we use the outcome data from treatments 2 through 7. The outcome data for treatment \( j \) is a cumulative distribution function \( F_j \). For each treatment \( j \), we use a kernel density estimator to construct the density \( \hat{f}_j \). Then, for each pair \( (A,B) \), we use these densities to construct an estimate of the function

\[
\hat{h}^{AB}(x) = \frac{\hat{f}^A(x) - \hat{f}^B(x)}{a^A - a^B}.
\]

For each triple \( (A,B,C) \), we then construct the predicted marginal incentives under contract \( C \) using data from contracts \( A \) and \( B \) according to:

\[
\hat{I}_{CA} = \int v(w_C(x)) \hat{h}^{AB}(x) \, dx.
\]

Using the estimates of the agent’s marginal incentives under contracts \( A \) and \( B \), we can then estimate the relevant parameters of the agent’s cost function:

\[
\hat{\varepsilon}^{AB} = \frac{\ln a^A - \ln a^B}{\ln \hat{I}_A^{AB} - \ln \hat{I}_B^{AB}}
\]

and \( \hat{\beta}^{AB} = \ln a^A - \hat{\varepsilon}^{AB} \ln \hat{I}_A^{AB} \). For this exercise, it does not matter which of the two contracts

---

6 For this exercise, we will not use data from treatment 1, the no-incentives treatment. Our baseline model predicts that under the contract \( w^1(x) = 100 \), subjects would exert zero effort. They do not. We discuss how to incorporate external incentives such as intrinsic motivation or boredom avoidance into our model, which is important for accounting for these types of results in Section 5.2.

7 We use the triweight kernel with the bandwidth determined by the Silverman Rule-of-Thumb. See Hansen (2009) for details. We ignore observations with \( x > 3500 \) following DellaVigna and Pope’s observation that it is physically impossible to achieve more than 3500 points during the 10-minute interval and it is likely that these individuals are using bots. The results are similar if we use a different kernel estimator or we incorporate all observations.
we suppose to be the status quo and test contracts. Finally, our predictions for average points accumulated in treatment $C$ are $\ln \hat{a}_{C}^{AB} = \hat{\beta}^{AB} + \hat{\varepsilon}^{AB} \ln \hat{I}_{C}^{AB}$.

We focus first on what we refer to as **homogeneous A/B tests**, that is, A/B tests in which treatments $A$ and $B$ are in the same class; i.e., they are both piece-rate treatments or both bonus treatments. We discuss **hybrid A/B tests**, where treatments $A$ and $B$ are not in the same class, at the end of this section. For homogeneous A/B tests, we will say that a prediction is a **within-class prediction** if treatments $A$, $B$, and $C$ are in the same class. We will say that a prediction is an **across-class prediction** if treatments $A$ and $B$ are in the same class, but treatment $C$ is in a different class.

The following result summarizes our main findings for homogeneous A/B tests.

**Result 1.** For homogeneous A/B tests,
(a) predicted out-of-sample performance is highly correlated with actual performance,
(b) predictions are close to actual performance for both within-class and across-class predictions, and
(c) predictions for a given treatment are similar no matter which pair of contracts is used to construct the prediction.

Before we discuss our results, it is important to note two features of our out-of-sample predictions. First, they are based on theory and are not fitted to the in-sample data. Second, they collapse information on the entire distribution of output to create a one-dimensional index of predicted incentives for each out-of-sample treatment. Systematic prediction errors therefore can provide us with useful information about the extent to which our underlying principal–agent model fails to capture important drivers of the agent’s behavior.

Figure 1 plots our predictions against the actual performance for each treatment for all homogeneous A/B tests. The horizontal axis, depicts the actual effort, $a_{C}$, for treatments $C \in \{2, 3, \ldots, 7\}$, while the vertical axis plots our prediction, $\hat{a}_{C}^{AB}$. Across all our predictions, the correlation between $\hat{a}_{C}^{AB}$ and $a_{C}$ is 0.94. Notice that all predictions are close to the 45-degree line, depicted by the dashed line, illustrating parts (a) and (b) of Result 1. Moreover, for any $C$, the predictions from all (homogeneous) A/B tests are similarly, illustrating part (c).

We also compute, for each triple $(A, B, C)$, the absolute percentage error (APE) of our

---

8That is because, by construction, these objects are symmetric in $(A, B)$: $\hat{g}^{AB} = \hat{g}^{BA}$, $\hat{\varepsilon}^{AB} = \hat{\varepsilon}^{BA}$, and $\hat{\beta}^{AB} = \hat{\beta}^{BA}$.
Figure 1: This figure plots our predictions against the actual performance for each treatment for all homogeneous A/B tests. The horizontal axis, depicts the actual average effort, $a_C$, for treatments $C \in \{2, 3, \ldots, 7\}$, while the vertical axis plots our prediction, $\hat{a}_C^{AB}$.

The average APE across all our predictions is 1.59 percent. As a comparison, average performance in treatment 7 is 20 percent higher than in treatment 2. We can break down these predictions by whether they are within class or across class. Across all within-class predictions in which treatments $A$, $B$, and $C$ are all piece-rate treatments, the average APE is only 0.66 percent. That is, A/B tests using piece-rate treatments accurately predict out-of-sample performance in piece-rate treatments.

Next, we can look at across-class predictions. For those predictions where $A$ and $B$ are bonus treatments, and $C$ is a piece-rate treatment, the average APE is 0.99 percent. The predictions are slightly worse for predictions where $A$ and $B$ are piece-rate treatments, and $C$ is a bonus treatment. There, the average APE is 2.71 percent, and as Figure 1 shows, they systematically underestimate performance. We discuss why at the end of this section.

Finally, Figure 1 also shows that the estimates of each treatment’s performance are tightly clustered. To quantify this result, we can compute, for each treatment $C$, the coefficient of variation of the predictions of $a_C$. The average coefficient of variation across the six
treatments is 0.7% and ranges between 0.21% for treatment 3 and 2% for treatment 2.

These results are summarized in Column 3 of Panel A of Table 2. This panel also shows two additional results. First, the worst-case APE is also small. This is true both for within-class predictions and across-class predictions. Second, the quality of predictions described in Result 1 is not sensitive to our assumptions about the agent’s coefficient of relative risk aversion. The prediction accuracy is also similar if the agent’s utility is assumed to belong to a different class of functions.\footnote{To be specific, if \( v(\omega) = 10^3 \omega - B \omega^2 \) and we vary \( B \) from zero to one (thus ensuring that marginal utility is always non-negative), the average APE varies between 1.44% and 1.76% for homogeneous A/B tests, and between 2.04% and 2.19% for hybrid A/B tests.}

<table>
<thead>
<tr>
<th>Table 2: Out-of-Sample Effort Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient of relative risk aversion</td>
</tr>
<tr>
<td>Panel A: Homogeneous A/B Tests</td>
</tr>
<tr>
<td>Corr ( (\hat{a}^{AB}_C, a_C) )</td>
</tr>
<tr>
<td>Average APE (%)</td>
</tr>
<tr>
<td>Within-class</td>
</tr>
<tr>
<td>Across-class: piece-rate predictions</td>
</tr>
<tr>
<td>Across-class: bonus predictions</td>
</tr>
<tr>
<td>Worst-case APE (%)</td>
</tr>
<tr>
<td>Within-class</td>
</tr>
<tr>
<td>Across-class: piece-rate predictions</td>
</tr>
<tr>
<td>Across-class: bonus predictions</td>
</tr>
<tr>
<td>Avg. CV of estimates (%)</td>
</tr>
<tr>
<td>Panel B: Hybrid A/B Tests</td>
</tr>
<tr>
<td>Corr ( (\hat{a}^{AB}_C, a_C) )</td>
</tr>
<tr>
<td>Average APE (%)</td>
</tr>
<tr>
<td>Worst-case APE (%)</td>
</tr>
<tr>
<td>Avg. CV of estimates (%)</td>
</tr>
</tbody>
</table>

Table 2: This table reports summary statistics for effort predictions under different assumptions for the agent’s (constant) coefficient of relative risk aversion. Panel A reports the correlation between predicted and actual effort, the average and worst-case absolute percentage error (APE), and the coefficient of variation of the estimates for homogeneous A/B tests. Panel B reports these quantities for the hybrid A/B tests.

While Result 1 and Panel A of Table 2 focus on homogeneous A/B tests, we can also carry out the same exercise for hybrid A/B tests. Panel B describes the performance for such tests. Across all predictions involving hybrid A/B tests, the correlation between \( \hat{a}^{AB}_C \) and \( a_C \) is 0.84, and the average APE is 2.16 percent. On average, hybrid A/B tests tend to perform almost as well as homogeneous A/B tests, but for some of the \( (A, B) \) pairs, they do much
worse: The worst-case APE across all hybrid tests is 10.7%. The hybrid \((A, B)\) pairs that perform relatively poorly are \((4, 6), (5, 6),\) and \((5, 7)\), and this poor performance reflects our implicit assumption that subjects choose effort once-and-for-all.

To see why, let us focus on the \((4, 6)\) pair—the lessons are similar when we look at \((5, 6)\) and \((5, 7)\). The output distributions under these two treatments have distinctly different patterns, as illustrated in Figure 4. In particular, for treatment 4, which is a piece-rate treatment, performance is roughly symmetrically distributed around the average. For the bonus treatment 6, however, performance spikes just over \(x = 2000\), the threshold for receiving the bonus. Given that the average performance in these two treatments are quite close to each other, with \(a_4 = 2132\) and \(a_6 = 2136\), because these averages appear in the denominator of (4), the estimated function \(\hat{h}_{AB}\) magnifies these differences. For, say, the \((2, 6)\) pair, we see similarly distinct patterns. Since the average performance in treatment 2, \(a_2 = 1883\), is significantly lower than in treatment 6, however, our out-of-sample predictions are less influenced by these patterns.

The reason why A/B tests comprising piece-rate treatments underestimate performance under the bonus treatments is related. To be specific, the function \(\hat{h}_{AB}\) estimated using bonus treatments takes very positive values for large \(x\), and very negative values for small \(x\). As a result, the marginal incentives generated by a contract which pays a lump-sum bonus if \(x \geq 2000\) are large. In contrast, the \(\hat{h}_{AB}\) estimated using piece-rate treatments takes more moderate values, and consequently, it under-predicts the marginal incentives, and hence the effort, generated by such a bonus contract.

5.2 Performance of Optimal Adjustments

For our second exercise, we will assess the empirical performance of our solution to the principal’s problem \((\text{Adj})\). To do so, we must first develop a benchmark to compare it against. For this, we will again use DellaVigna and Pope’s (2017) data and will proceed in two steps. First, we will build a benchmark model using the data from several of the treatments. For each treatment, we will take the associated contract, \(w^C\), to be the status quo contract, and we will compute the optimal contract \(w^* (w^C)\) that solves the principal’s problem using the parameters from this benchmark model and gives the agent at least as much expected utility as \(w^C\).

Second, for each pair of contracts \((A, B)\) belonging to the same class, we will take the information from the A/B test involving these two contracts, and we will compute the optimally adjusted contract \(w^{AB} (w^C)\) that solves \((\text{Adj})\) and gives the agent at least as much expected utility as \(w^C\). We will then compare the performance of \(w^* (w^C)\) and \(w^{AB} (w^C)\) as evaluated under the benchmark model.
5.2.1 The Benchmark Model and Optimal Adjustments

We will now describe in detail how we construct our benchmark model. Throughout, we will use tildes to denote components of the benchmark model. First, we estimate the output distribution $\tilde{f}(x|a)$ and its derivative with respect to $a$, $\tilde{f}_a(x|a)$ for all $x \in [0, 3500]$ and for all $a$ within a range, which we will describe below.\footnote{As explained in Section 5.1 we use the tricube kernel with the bandwidth determined by the Silverman Rule-of-Thumb, and have excluded observations with $x > 3500$.} Next, assuming that Condition 2 holds, we estimate the parameters of the agent’s cost function. As in the previous section, we will assume that the agent has constant-relative-risk-aversion (CRRA) preferences over money, so that $\tilde{v}'(\omega) = \omega - \tilde{\rho}$, and we will assume that $\tilde{\rho} = 0.3$ and assess the sensitivity of our results to this assumption. Finally, we will also need to make an assumption about the principal’s gross profit margin $\tilde{m}$. In particular, we will assume that $\tilde{m} = 0.2$. We discuss this choice below.

Output Distribution To construct the output distribution $\tilde{f}(x|a)$, we proceed in two steps. First, we use outcome data for treatments 1 to 5—the no-incentives treatment and the piece-rate treatments. These outcome data are a set of cumulative distribution functions $F(x|a_C)$, one for each of the five treatments $C \in \{1, \ldots, 5\}$. As discussed in the previous section, we use a kernel density estimator to construct the density $\hat{f}(x|a_C)$ for each treatment $C \in \{1, \ldots, 5\}$. We assume that $\tilde{f}(x|a) = \hat{f}(x|a_{C})$ for all $a \in \{a^1, \ldots, a^5\}$, and for each $x$, we use a spline interpolation to estimate $\tilde{f}(x|a)$ for other values of $a$ between $a^1$ and an upper bound, $\bar{a}$. The spline interpolation is not guaranteed to satisfy $\tilde{f}(x|a) \geq 0$ for all $x$ for choices of $a$ outside the bounds of our data. We chose our upper bound $\bar{a}$ to be 2187, which is the largest value $\bar{a}$ such that $\tilde{f}(x|a) \geq 0$ for all $a \in [a^1, \bar{a}]$ for all $x$. Given our output distribution function $\tilde{f}(x|a)$, we approximate $\tilde{f}_a(x|a) = \tilde{f}(x|a+1) - \tilde{f}(x|a)$. We discuss the choice of this particular subset of treatments for constructing $\tilde{f}(x|a)$ in footnote \footnote{As explained in Section 5.1 we use the tricube kernel with the bandwidth determined by the Silverman Rule-of-Thumb, and have excluded observations with $x > 3500$.}

Agent’s Cost Function We first return to an issue that came up in the previous section. The contract associated with treatment 1 provides no marginal incentives: It is given by $w^1(x) = 100$ for all $x$. The baseline model would therefore predict zero effort. Yet subjects in treatment 1 scored 1521 points on average. To rationalize the fact that subjects chose strictly positive effort levels in this treatment, we modify Condition 2 and assume that the agent’s cost function is given by $\tilde{c}'(a) = e^{-\tilde{\beta}/\tilde{\epsilon}} a^{1/\tilde{\epsilon}} - \tilde{I}_0$ for some $\tilde{I}_0 \geq 0$. This parameter can be interpreted as the agent’s external incentives: they may come from intrinsic motivation, longer-term career incentives, or in the case of this experiment, the fact that it may be fun to challenge yourself to see how many points you can score. Constructing the agent’s cost
function therefore requires estimating three parameters from the data: $\tilde{\epsilon}$, $\tilde{\beta}$, and $\tilde{I}_0$. Table 3 reports the estimates for these parameters using nonlinear least squares estimation.\footnotemark

| Table 3: Fitted parameters for the benchmark model |
|-----------------|-----------------|-----------------|
| $\tilde{\epsilon}$ | $\tilde{\beta}$ | $\tilde{I}_0$ |
| 0.0322          | 7.8184          | $6.528 \times 10^{-7}$ |

Table 3: This table displays the fitted parameters for the benchmark model. They are computed using data from treatments 1 to 5 and using a nonlinear least squares estimation procedure.

**Benchmark Optimal Contract** We then solve for the principal’s optimal contract in our benchmark model. Recall that the optimal contract depends on what the status quo contract is because it determines the utility that the principal must provide to the agent. We therefore compute an optimal contract for each treatment $C \in \{2, \ldots, 7\}$. We take $w^C$ to be the status quo contract, and we solve for the principal’s optimal contract, $w^* \left( w^C \right)$, by solving the following two-step problem.

First, for each $a \in [a^1, \pi]$, using the CVX software for Matlab (Grant and Boyd, 2013), we find the cost-minimizing contract that solves

$$
\tilde{C} \left( a; w^C \right) = \min_{w(\cdot)} \int w(\cdot) \tilde{f} (x|a(w)) \, dx
$$

subject to the constraint that effort $a$ is incentive compatible,

$$
\int \tilde{v} \left( w(\cdot) \right) \tilde{f} \left( x|a \right) \, dx - \tilde{c} \left( a \right) \geq \int \tilde{v} \left( w(\cdot) \right) \tilde{f} \left( x|a' \right) \, dx - \tilde{c} \left( a' \right) \quad \text{for all } a',
$$

the constraint that the agent is at least as well off as under the status quo contract

$$
\int \tilde{v} \left( w(\cdot) \right) \tilde{f} \left( x|a \right) \, dx - \tilde{c} \left( a \right) \geq \int \tilde{v} \left( w^C(\cdot) \right) \tilde{f} \left( x|a \left( w^C \right) \right) \, dx - \tilde{c} \left( a \left( w^C \right) \right),
$$

and two additional constraints.\footnotemark First, we impose the constraint that $w(\cdot) \geq 100$ for all $x$

\footnotetext{11}{To be specific, for each treatment $C$, we compute $\tilde{f}^C = \int v(\cdot) \tilde{f}^C (x|a^C) dx$, and minimize $\sum_{C=1}^{7} \left[ \log(a_i) - \beta - \epsilon \log(\tilde{f}^C + I_0) \right]^2$ to obtain $\tilde{\epsilon}, \tilde{\beta},$ and $\tilde{I}_0$. Notice that to estimate three parameters, one needs outcome data from at least three treatments. As a result, we do not consider A/B tests that include the no-incentives treatment, $C = 1$. Recall that we construct $\tilde{f} (x|a)$ using outcome data from only treatments 1 to 5. This subset minimizes the objective in the above optimization program.

\footnotetext{12}{Since the first-order approach is generically not valid, we impose a global incentive compatibility constraint, requiring that the target effort level gives the agent a larger expected utility than any other (possibly non-local) deviation.}
to capture the fact that each subject was paid a $1 participation fee. Second, we impose the constraint that \( w(x) \) is weakly increasing in \( x \).

For the second step, we do a line search to solve for the principal’s optimal choice of \( a \)

\[
\pi^* (w^C) = \max_{a \in [a^1, \bar{a}]} \bar{m} a - \tilde{C} (a; w^C).
\]

Denote the optimal effort choice by \( a^* (w^C) \). Solving this problem gives us three objects that we use as our benchmark. It gives us the principal’s optimal expected profits \( \pi^* (w^C) \), the optimal effort level she implements, \( a^* (w^C) \), and the optimal contract she puts in place to implement that effort level, \( w^* (w^C) \).

We conclude this section with a brief discussion of our choice of \( \bar{m} \). Our goal was two-fold. We wanted to choose a value of \( \bar{m} \) that is high enough so that none of the status quo contracts yield negative profits, that is, \( \pi (w^C) \geq 0 \) for \( j \in \{2, \ldots, 7\} \). And we wanted to choose a value that is low enough so that the optimal effort choice \( a^* (w^C) \) is below \( \bar{a} \) for most treatments. Our choice of \( \bar{m} = 0.2 \) satisfies these two conditions. We also show in Table 4 how the main pattern of results varies with \( \bar{m} \in [0.15, 0.25] \).

**Optimal Adjustments Given an A/B Test** We then solve for the principal’s optimal adjustment to each contract given information from an A/B test. Again, for each treatment \( C \in \{2, \ldots, 7\} \), we take \( w^C \) to be the status quo contract. For each pair \( (A, B) \), we construct an output probability density function and an agent cost function using the outcome data from contracts \( w^A \) and \( w^B \). In particular, we construct \( \hat{g}^{AB} \) and \( \hat{h}^{AB} \) as in the previous section. From these two functions, we construct an output probability density function \( \hat{f}^{AB} \) that satisfies \( \hat{f}^{AB} (x | a) = \hat{g}^{AB} (x) + a \hat{h}^{AB} (x) \) for all \( x \) and for all \( a \in [a^1, \bar{a}^{AB}] \), where \( \bar{a}^{AB} \) is chosen so that \( \hat{f}^{AB} (x | a) \geq 0 \) for all \( x \) and for all \( a \leq \bar{a}^{AB} \). The cost-function parameters \( \hat{\varepsilon}^{AB} \) and \( \hat{\beta}^{AB} \) are constructed as in the previous section, so that the agent’s cost function is \( \hat{c}^{AB} (x) = e^{-\hat{\varepsilon}^{AB}/\hat{\beta}^{AB}} a^{1/\hat{\varepsilon}^{AB}} \). We again assume that the agent has constant-relative-risk-aversion preferences over money, \( \hat{v}' (\omega) = \omega^{-\hat{\beta}} \), and we assume that \( \hat{\beta} = 0.3 \).

We then solve for the principal’s optimal adjustment by solving the following two-step problem. First, for each \( a \in [a^1, \bar{a}^{AB}] \), we solve for the cost-minimizing contract that solves

\[
\hat{C}^{AB} (a; w^C) = \min_{w(\cdot)} \int w(x) \hat{f}^{AB} (x | a(w)) \, dx
\]

\( \text{This constraint serves two purposes. First, without it, the optimal effort is always equal to the upper bound, } \bar{a}, \text{ which implies that any optimal adjustment will mechanically implement an effort that is weakly smaller than optimal—an undesirable handicap to the algorithm. Second, the optimal contract need not be monotone, which can motivate gaming and other undesirable behaviors. See also Innes (1990) and Oyer (2000) for a motivation of this constraint.} \)
subject to the agent’s first-order condition for effort

\[ \hat{c}^{AB} (a) = \int \hat{v} (w (x)) \hat{h}^{AB} (x) \, dx, \]

the constraint that the agent is at least as well off as under the status quo contract

\[ \int \hat{v} (w (x)) \hat{f}^{AB} (x \mid a) \, dx - \hat{c}^{AB} (a) \geq \int \hat{v} (w^{C} (x)) \hat{f}^{AB} (x \mid a (w^{C})) \, dx - \hat{c}^{AB} (a (w^{C})) \]

as well as the two additional constraints we imposed in the benchmark optimal contract exercise: \( w (x) \geq 100 \) for all \( x \) and \( w (x) \) is weakly increasing in \( x \).

For the second step, we do a line search to solve for the principal’s optimal choice of \( a \)

\[ \max_{a \in [a_1, \pi^{AB}]} \tilde{m} a - \hat{C}^{AB} (a; w^{C}), \]

and let \( \pi^{AB} (w^{C}) \) be the resulting profits the principal receives under the benchmark model evaluated at the contract that solves this problem. Solving this problem gives us three objects that we compare to our benchmark. It gives us the principal’s optimally adjusted profits, \( \pi^{AB} (w^{C}) \), the optimal effort level she implements, \( a^{AB} (w^{C}) \), and the optimal contract she puts in place to implement that effort level, \( w^{AB} (w^{C}) \).

5.2.2 Performance

We will now discuss the performance of optimal adjustments. To do so, we first have to define what it means for optimal adjustments to perform well. In particular, we want to compare the principal’s expected profits under the optimal adjustment to the profits she would attain if she put in place the optimal contract, starting from some status quo contract. We will take the status quo contracts to be the contracts associated with treatments 2 through 7. The performance comparison is therefore going to depend on which treatment we are looking at, as well as which pair of contracts we use for our A/B test.

Formally, let us define two quantities for each treatment \( C \). First, we will define the maximum gains for treatment \( C \) to be the maximized gains relative to the status quo contract, that is,

\[ \text{MaxGains}^{C} = \pi^{*} (w^{C}) - \pi (w^{C}), \]

\(^{14}\)An implication of Condition 2 is that the first-order approach is valid. Therefore, it is without loss of generality to replace the agent’s incentive compatibility constraint with the corresponding first-order condition.

\(^{15}\)Both the principal’s profits and the effort level that she implements are evaluated according to the benchmark model we estimated in the beginning of this section.
where \( \pi(w^C) \) is the expected profits under status quo contract \( w^C \). Second, we will define the **average realized gains for treatment** \( C \) to be the average maximized profits under status quo contract \( w^C \) across all homogeneous A/B tests, that is,

\[
\text{AvgGains}^C = \frac{1}{|\text{Hom}|} \sum_{A,B \in \text{Hom}} \pi^{AB}(w^C) - \pi(w^C),
\]

where \( \text{Hom} \equiv \{(A, B) \mid (A, B) \text{ is a homogeneous pair}\} \). The following result summarizes our main findings for the performance of optimal adjustments.

**Result 2.** For homogeneous A/B tests,

(a) the average realized gains are about 68% of the maximum gains across all treatments,

(b) optimally adjusted efforts are close (but not equal) to optimal efforts, and

(c) optimally adjusted contracts tend not to be cost minimizing for the effort levels they induce.

Figure 2 illustrates the first part of Result 2. The horizontal axis plots maximum gains, while the vertical axis plots realized gains. Each pentagram represents for treatment \( C \), the point with coordinates \((\text{MaxGains}^C, \text{AvgGains}^C)\).

The first part of Result 2 shows that optimal adjustments perform quite well. It is a comparison of the average realized gains across all treatments to the maximum gains across all treatments. In particular, the quantity \((1/6) \sum_{C=2}^7 \text{AvgGains}^C\) is \(7.14\), about 68% of the quantity \((1/6) \sum_{C=2}^7 \text{MaxGains}^C\), which is \(10.55\). Notice that the approximately $3 gap between average and maximum gains has little variance across treatments, as illustrated by the ordinary least squares fitted line whose slope and intercept is close to 1 and −3, respectively.

The last two parts of Result 2 assess the dimensions along which optimal adjustments perform better and worse. Optimal adjustments tend to do well in terms of motivating an effort level that is close to the optimal one. But they tend to do so at higher cost than necessary. Figure 3 below illustrates the second part of Result 2. On the horizontal axis, it plots the optimal effort change for each treatment, \( a^*(w^C) - a^C \). On the vertical axis, it plots the average realized effort adjustment across all homogeneous A/B tests, that is, \((1/|\text{Hom}|) \sum_{A,B \in \text{Hom}} a^{AB}(w^C) - a^C\).

This figure illustrates several points. First, the optimal effort change varies quite a bit across treatments. For treatments 5 and 7, the optimal effort change is negative, and for treatment 2, it is almost 200. Second, the average realized effort adjustment is pretty close to the optimal effort change; i.e., \((1/|\text{Hom}|) \sum_{A,B \in \text{Hom}} a^{AB}(w^C) - a^C\) is pretty close to the
Figure 2: This figure compares the average gains to the maximum gains. In particular, each pentagram represents for treatment $C$, the point with coordinates $(\text{MaxGains}^C, \text{AvgGains}^C)$. By construction, the average realized gains lie below the 45-degree line, depicted by the dashed red line. The green dotted line represents the ordinary least squares fit through the pentagrams.

45-degree line. Averaging across all six treatments, the average effort deviation is $-7.12$, suggesting that on average, optimal adjustments induce 7.12 fewer points than would be optimal in the benchmark model. Given that each point yields $\tilde{m} = 0.2$ dollars in profits for the principal, on average, the principal is losing about $1.42$ in revenues from implementing too low of an effort level, or approximately two-fifths of the gap between the average and maximum gains.

Next, we compare two quantities for each treatment $C$. For each pair $(A, B)$, the optimally adjusted contract $w^{AB}(w^C)$ induces effort level $a^{AB}(w^C)$ and therefore costs the principal

$$\text{WageBill}^{AB}(w^C) \equiv \int w^{AB}(w^C)(x) \tilde{f}(x|a^{AB}(w^C)) \, dx.$$ 

We want to compare this wage bill to the cost of the cheapest contract that implements the same effort level, which is given by $\hat{C}(a^{AB}(w^C); w^C)$. For each treatment $C$, let us define the average overpayment to be $(1/|\text{Hom}|) \sum_{A,B \in \text{Hom}} \text{WageBill}^{AB}(w^C) - \hat{C}(a^{AB}(w^C); w^C)$. Across the six treatments, the average overpayment is about $1.79$. 

27
Figure 3: This figure compares the effort implemented by the optimal adjustment using an A/B test to the optimal effort. To be specific, for each treatment $C$, each pentagram represents the point with x-coordinate $a^* (w^C) - a^C$ and y-coordinate $(1/|Hom|) \sum_{A,B \in Hom} a^{AB} (w^C) - a^C$. The dashed red line is the 45-degree line.

Table 4 reports these summary statistics for different values of $\tilde{\rho}$, the coefficient of CRRA we used to estimate the benchmark model, $\hat{\rho}$, the coefficient of CRRA that the principal assumed to find the optimally adjusted contract given an A/B test, and $\tilde{m}$, the principal’s profit margin. Observe that both average and maximum gains increase with $\tilde{m}$, but the gains ratio has little variance. All summary statistics are relatively insensitive to the values of $\tilde{\rho}$ and $\hat{\rho}$.

Recall that our objective is to find, given an A/B test and a benchmark treatment $C$, the profit-maximizing contract that gives the agent at least as much utility as $w^C$. We conclude this section by discussing the performance of our algorithm on this dimension. To do so, we will compare the agent’s expected utility under the optimally adjusted contract, $U^{AB}(w^C) = \int v(w^{AB}(w^C)(x)) \tilde{f}(x|a^{AB}(w^C))dx - \tilde{c}(a^{AB}(w^C))$, to his expected utility under the benchmark contract $w^C$, $U(w^C) = \int v(w^C(x))\tilde{f}(x|a(w^C))dx - \tilde{c}(a(w^C))$. So the aim is for the ratio $U^{AB}(w^C)/U(w^C)$ to be greater than one. Figure 5 illustrates its cumulative

---

16Notice that the average loss due to implementing a suboptimal effort, $\tilde{m} \times \text{Average Effort Deviation}$, and the average overpayment do not add up to the difference between the maximum and average gains. This is because the overpayment is defined as the difference between the wage bill of the optimally adjusted contract given an A/B test and the cost-minimizing contract that implements the same effort, which of course, need not equal the optimal effort.
Table 4: Performance of Optimal Adjustments and Sensitivity Analysis

<table>
<thead>
<tr>
<th></th>
<th>Model coeff. of CRRA ($\hat{\rho}$)</th>
<th>Test coeff. of CRRA ($\hat{\rho}$)</th>
<th>Profit margin ($\hat{m}$)</th>
<th>Average Gains ($)</th>
<th>Maximum Gains ($)</th>
<th>Gains Ratio (%)</th>
<th>Average Effort Deviation</th>
<th>Average Overpayment ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.15</td>
<td>0.25</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Average Gains ($)</td>
<td>7.14</td>
<td>5.17</td>
<td>9.39</td>
<td>7.22</td>
<td>7.12</td>
<td>6.80</td>
<td>7.34</td>
<td>10.55</td>
</tr>
<tr>
<td>Gains Ratio (%)</td>
<td>67.71</td>
<td>67.88</td>
<td>69.47</td>
<td>67.24</td>
<td>67.23</td>
<td>64.51</td>
<td>69.61</td>
<td>67.88</td>
</tr>
<tr>
<td>Average Effort Deviation</td>
<td>-7.12</td>
<td>-7.74</td>
<td>-6.21</td>
<td>-8.08</td>
<td>-6.95</td>
<td>-8.73</td>
<td>-6.05</td>
<td>1.79</td>
</tr>
<tr>
<td>Average Overpayment ($)</td>
<td>1.79</td>
<td>1.31</td>
<td>2.09</td>
<td>1.77</td>
<td>1.699</td>
<td>2.13</td>
<td>1.69</td>
<td>2.13</td>
</tr>
</tbody>
</table>

Table 4: This table reports for different values of the parameters $\hat{\rho}$, $\hat{\rho}$, and $\hat{m}$, the average and maximum gains, the gains ratio, the average effort deviation, and the average overpayment, averaged across $C \in \{2, ..., 7\}$.

distribution function across all homogeneous A/B tests: it is equal to 1.01 on average, and it is larger than one in 62 out of the 84 cases.

6 Beyond the Classic Model

In this section, we show how to extend our analysis in Section 3 in three directions. First, we consider settings in which the agent’s effort and output are multidimensional. Effort-substitution patterns become important for optimal adjustments, but we show that they can be identified with additional test contracts. Second, we consider settings in which the principal is restricted to offering contracts from a particular parametric class, for example, linear contracts or bonus contracts. We show that our main sufficient statistic result continues to hold in these settings. Finally, we show how to incorporate worker heterogeneity.

6.1 Multi-dimensional Effort

In this section, we extend our main model to the case where the agent’s effort and output are multidimensional. As an example, suppose the agent can exert effort towards selling $M$ different products to his customers. Suppose that for each product $i \in \{1, \ldots, M\}$, he chooses effort $a_i$, which in turn generates sales $x_i$. Assume that $x \sim f(\cdot | a)$, where $a$ and $x$ are vectors of efforts and sales, and for simplicity, assume that $x_i$ and $x_j$ are independent conditional on $a$. We will normalize each effort so that $a_i = \mathbb{E} [x_i | a_i]$. The agent is paid according to a contract $w(x)$, and the cost of choosing effort vector $a$ is $c(a)$, where $c$ satisfies the usual conditions.
Figure 4: This figure illustrates the estimated probability density function for treatments 4 and 6.

Figure 5: This figure illustrates the cumulative distribution function of the ratio $U^{AB}(w^C)/U(w^C)$.

The agent’s utility given contract $w$ is

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

where the integral is taken with respect to the entire vector $x$. Assuming the first-order approach is valid, we can use the same approach as in Section 3 to derive how the agent’s utility and effort respond to a local adjustment of the contract $w$ in the direction $t$. In particular,

$$D u(w, t) = \int tv'(w) fdx,$$

and, for each $i \in \{1,\ldots,M\}$,

$$\sum_{k=1}^{M} \left[ c_{i,k} - \int v(w) f_{i,k} dx \right] DA_{k}(w, t) = \int tv(w) f_{i} dx,$$

where $f_{i}(x|a) = \partial f(x|a)/\partial a_{i}$, $c_{i,k}(a) \equiv \partial^{2}c(a)/\partial a_{i}\partial a_{k}$ and similarly for $f_{i,k}$. We have dropped the dependence of these functions on $x$ and $a$ to simplify the expressions.

Given contract $w$, let us define the agent’s Hessian matrix $A$ to be an $M \times M$ symmetric matrix with elements $A_{i,k} = c_{i,k} - \int v(w) f_{i,k} dx$. Let us also define the agent’s marginal-
incentives matrix under adjustment \( t \) to be the \( M \times 1 \) matrix \( B (t) \) with elements \( B_i(t) = \int tv(w) f_i dx \). We can then rewrite (2) as \( Da(w, t) = A^{-1} B(t) \), where \( Da(w, t) \) denotes the \( M \times 1 \) matrix with \( k^{th} \) element \( Da_k(w, t) \).

Next, we turn to the principal’s profits. Using the same approach as in Section 3, adjusting a contract \( w \) in the direction \( t \) changes her profit according to the differential

\[
\mathcal{D} \pi(w, t) = \sum_{i=1}^{M} \left[ m - \int w(x) f_i(x|a(w)) \, dx \right] Da_i(w, t) - \int t(x) f(x|a(w)) \, dx.
\]

Given a status quo contract, \( w^A \), the principal solves

\[
\max_{||t|| \leq 1} \mathcal{D} \pi(w^A, t) \quad \text{subject to} \quad \mathcal{D} u(w^A, t) \geq 0.
\]

Recall that Proposition 1 shows that, for the \( M = 1 \) case, a local A/B test reveals \( Da(w, t) \) for a particular adjustment, \( t \), and enables one to compute how the agent’s marginal incentives change for any \( t \). By the same logic, when \( M \geq 2 \), a single local A/B test reveals the vector \( Da(w, t) \) as well as the vector of functions \( (f_1, \ldots, f_M) \). This information suffices for constructing the agent’s marginal incentives matrix \( B(t) \) for any \( t \).

When \( M = 1 \), the agent’s Hessian matrix \( A \) is a singleton, and Proposition 1 showed that it can be identified with a single local A/B test. When \( M \geq 2 \), the agent’s Hessian matrix contains \( M (M + 1) / 2 \) elements, which cannot all be inferred from a single local A/B test. Since each local A/B test generates \( M \) equations, \( [(M + 1) / 2] \) local A/B tests are required to infer the agent’s Hessian matrix and hence be able to evaluate \( Da(w, t) \) and \( \mathcal{D} \pi(w, t) \) for any \( t \).

### 6.2 Parametric Classes of Contracts

Firms often restrict attention to a particular class of contracts. For instance, linear contracts are very common, as are piece-wise linear and bonus contracts. Such contracts may be chosen due to their simplicity, or because of considerations outside of our model. For example, Carroll (2015) and Barron, Georgiadis, and Swinkels (2020) provide rationales for linear contracts, Innes (1990) for piece-wise linear contracts, and Oyer (2000), and Georgiadis and Szentes (2020) for bonus contracts.

In this section, we address the same questions as in Section 3 assuming that the principal only considers contracts which belong to a particular parametric class, denoted by \( w_\alpha \), where \( \alpha \) is a finite vector of parameters. For example, if the principal considers only linear contracts, then \( w_\alpha(x) = \alpha_1 + \alpha_2 x \), and she chooses the parameters \( \alpha_1 \) and \( \alpha_2 \). If she restricts attention
to bonus contracts, then $w_\alpha(x) = \alpha_1 + \alpha_2 I_{x \geq \alpha_3}$ for some $\alpha_1$, $\alpha_2$, and $\alpha_3$ that remain to be chosen. We maintain the assumptions imposed in Section 2, with the exception that both the status quo and the test contract belong to the particular parametric class, with parameters $\alpha^A$ and $\alpha^B$, respectively.

Given an arbitrary parameter and adjustment vector, $\alpha$ and $z$, respectively, we will consider contracts of the form $w_{\alpha + \theta z}$. Following the analysis in Sections 3, we can evaluate how the agent’s utility and effort changes in response to a local adjustment to the vector $\alpha$ in the direction $z$, which yields

$$D u(w_\alpha, z) = \sum_i z_i \frac{\partial}{\partial \alpha_i} \int v(w_\alpha(x)) f(x|a(w_\alpha)) dx,$$

and

$$D a(w_\alpha, z) = \sum_i z_i \frac{\partial}{\partial \alpha_i} \int v(w_\alpha(x)) f_a(x|a(w_\alpha)) dx$$

Similarly, this adjustment changes the principal’s profit according to the differential

$$D \pi(w_\alpha, z) = \left[ m - \int w_\alpha(x)f_a(x|a(w_\alpha))dx \right] D a(w_\alpha, z) - \sum_i z_i \frac{\partial}{\partial \alpha_i} \int w_\alpha(x)f(x|a(w_\alpha))dx.$$

Given a status quo contract with parameters $w_{\alpha^A}$, the principal solves

$$\max_{z:||z|| \leq 1} D \pi(w_{\alpha^A}, z) \quad \text{subject to} \quad D u(w_{\alpha^A}, z) \geq 0.$$ 

Both the objective and the constraint are linear in $z$, implying that this is a standard convex program. Second, it is straightforward to adapt the analysis of Section 4 to consider non-local adjustments to the status quo contract. In general, this problem is not convex, so standard optimization techniques are not guaranteed to achieve the global maximum. However, as long as the dimension of $\alpha$ is sufficiently small, using grid-search is a feasible alternative.

### 6.3 Heterogeneous Workers

The goal of this section is to extend the methodology developed in Section 3 to the case in which the principal offers a common contract to a set of heterogeneous agent. We will argue that a local A/B test, together with knowledge about the agent heterogeneity suffices to find the optimal adjustment.

Suppose that each agent is indexed by a type, $\phi$, and agents with different types have different effort cost functions but are otherwise identical. We denote the cost of a type-$\phi$ agent choosing effort $a$ (which generates expected output $a$) by $c(a, \phi)$. Moreover, we assume
that the principal can classify agents into types, and we let $p^\phi$ denote the fraction of type-$\phi$ agents. In practice, this can be achieved, for example, if the principal observes time-series output data for each agent under one contract. A local A/B test enables the principal to estimate for each $\phi$, the density $f^\phi(\cdot|a^\phi(w))$ and the score $f^\phi_\phi(\cdot|a^\phi(w))$, and evaluate the differential $D^\phi(a(w,t))$ for any adjustment $t$.

Replicated the steps of the analysis in Section 3, the principal’s problem can be expressed as

$$\max_{t:||t|| \leq 1} \sum_\phi p^\phi \left[ m - \int w(x) f^\phi_s(x|a^\phi(w)) \, dx \right] D^\phi(a(w,t)) - \int t(x) f^\phi_s(x|a^\phi(w)) \, dx,$$

s.t. $\int t(x) v'(w(x)) f^\phi_s(x|a^\phi(w)) \, dx \geq 0$ for each $\phi$. (5)

It is well known that the design of performance pay may be used to induce selection, that is, attract more productive workers and induce less productive workers to exit (Lazear, 2000). To do so in our setting, the principal may restrict attention to perturbations that give at least as much utility as $\hat{w}$ to only a subset of the more productive types. Formally, this would imply that (5) must hold only for the types that the principal wants to attract, and must be violated for the types that she wants to repel from the firm or dissuade from joining. A full analysis of the selection effects associated with performance pay is beyond the scope of this paper, and is left for future work.

- Explain that it is straightforward to extend the analysis of Section 4 to heterogeneous workers.

7 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.

\footnote{We use the superscript $\phi$ to denote type-$\phi$ dependent quantities.}
References


A Proofs

Proof of Lemma 1.

Fix arbitrary u.s.c functions $w$ and $t$, and consider the contract $w + \theta t$ for some $\theta > 0$. By assumption, the agent’s effort satisfies the first-order condition

$$\int v(w(x) + \theta t(x)) f_a(x|a(w + \theta t)) dx = c'(a(w + \theta t)).$$

Differentiating this equation with respect to $\theta$ and taking the limit as $\theta \to 0$ yields

$$\int t(x)v'(w(x)) f_a(x|a(w)) = \left[ c''(a(w)) - \int v(w(x)) f_a(x|a(w)) \right] D_a(w, t),$$

and using the definition $D_I(w, t) := \int t(x)v'(w(x)) f_a(x|a(w))$, we obtain the desired expression for $D_a(w, t)$.

Next, consider the agent’s expected utility. Faced with contract $w + \theta t$, the agent’s expected utility is

$$u(w + \theta t) = \int v(w(x) + \theta t(x)) f(x|a(w + \theta t)) dx - c(a(w + \theta t)).$$

Differentiating with respect to $\theta$ and taking the limit as $\theta \to 0$ yields

$$D_u(w, t) = \int t(x)v'(w(x)) f(x|a(w)) dx + \left[ \int v(w(x)) f_a(x|a(w)) - c'(a(w)) \right] D_a(w, t)$$

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

where the second equality follows because the term in brackets is equal to 0 by the agent’s first-order condition.

Proof of Lemma 2.

This lemma follows immediately from the fact that for any $t$, $D_\pi(w, t)$ and $D_u(w, t)$ depend only on $f(x|a(w))$, $f_a(x|a(w))$, and $D_a(w, t)$, and not other parameters of the production environment.

Proof of Proposition 1.
Note that
\[
LAB \left( w^A, w^B \mid P \right) = \left( f^A, f^A_a, Da \left( w^A, w^B \mid P \right) \right)
\]
\[
LAB \left( w^A, w^B \mid \tilde{P} \right) = \left( \tilde{f}^A, \tilde{f}^A_a, Da \left( w^A, w^B \mid \tilde{P} \right) \right).
\]

If the first statement is true, then the second is obviously true. Next, suppose
\[
LAB \left( w^A, w^B \mid P \right) = LAB \left( w^A, w^B \mid \tilde{P} \right).
\]
It is immediate that \( f^A = \tilde{f}^A \) and \( f^A_a = \tilde{f}^A_a \). Next, note that for all \( t \),
\[
\mathcal{D}I \left( w^A, t \mid P \right) = \int tv' \left( w^A \right) f_a = \int tv' \left( w^A \right) \tilde{f}_a = \mathcal{D}I \left( w^A, t \mid \tilde{P} \right).
\]
By Lemma 1,
\[
\mathcal{D}a \left( w^A, t \mid P \right) = \frac{\mathcal{D}a \left( w^A, w^B \mid P \right)}{\mathcal{D}I \left( w^A, w^B \mid P \right)} \mathcal{D}I \left( w^A, t \mid P \right)
\]
\[
\mathcal{D}a \left( w^A, t \mid \tilde{P} \right) = \frac{\mathcal{D}a \left( w^A, w^B \mid \tilde{P} \right)}{\mathcal{D}I \left( w^A, w^B \mid \tilde{P} \right)} \mathcal{D}I \left( w^A, t \mid \tilde{P} \right),
\]
so \( \mathcal{D}a \left( w^A, t \mid P \right) = \mathcal{D}a \left( w^A, t \mid \tilde{P} \right) \) for all \( t \) if and only if \( \mathcal{D}a \left( w^A, w^B \mid P \right) = \mathcal{D}a \left( w^A, w^B \mid \tilde{P} \right) \),
which is true by supposition. \( \square \)

Proof of Proposition 2.

The optimization program given in \((Adj\text{local})\) can be rewritten as follows:

\[
\begin{align*}
\max_{\mu} & \quad \mu^* \int tv' \left( w^A \right) f^A_a dx - \int t f^A dx \\
\text{s.t.} & \quad \int tv' \left( w^A \right) f^A dx \geq 0 \\
& \quad \int t^2 dx \leq 1
\end{align*}
\]

where
\[
\mu^* := \frac{\left( m - \int w_A \tilde{f}_a dx \right) Da(w^A, w^B)}{\int (w^B - w^A)v'(w^A)f^A_a dx},
\]
and we have used that for any \( t \), \( Da(w^A, t) = Da(w^A, w^B) \int tv' \left( w^A \right) f^A_a dx \int (w^B - w^A)v'(w^A)f^A_a dx \) by Lemma 1. Let \( \lambda \geq 0 \) and \( \nu \geq 0 \) denote the dual multipliers associated with the first and
the second constraint. The Lagrangian

\[ L(\lambda, \nu) = \max_t \left\{ \nu + \int \left[ t \left( \lambda v'(w^A) f^A + \mu^* v'(w^A) f_{a}^A - \hat{f} \right) - \nu t^2 \right] dx \right\}. \tag{6} \]

For any \( \lambda \geq 0 \) and \( \nu > 0 \), we can optimize the integrand with respect to \( t \) pointwise. Noting that the integrand is differentiable with respect to \( t \), the corresponding first-order condition implies that

\[ t_{\lambda, \nu} = \frac{(\lambda f^A + \mu^* f_{a}^A) v'(w^A) - f^A}{2\nu}, \]

where \( t, f^A, f_{a}^A \), and \( w^A \) are functions of \( x \).\(^{18}\)

Next, we pin down the optimal \( \lambda \) and \( \nu \), by solving the following dual problem:

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

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

\[ \lambda^* = \max \left\{ 0, \frac{\int (f^A - \mu^* v'(w^A) f_{a}^A) v'(w^A) f_{a}^A dx}{\int (v'(w^A) f^A)^2 dx} \right\} \tag{7} \]

and

\[ \nu^* = \frac{1}{2} \sqrt{\int [(\lambda^* f^A + \mu^* f_{a}^A) v'(w^A) - f^A]^2 dx}. \tag{8} \]

Thus, the optimal perturbation,

\[ t^* = t_{\lambda^*, \nu^*} = \frac{(\lambda^* f^A(x) + \mu^* f_{a}^A(x)) v'(w^A(x)) - f^A(x)}{\sqrt{\int [(\lambda^* f^A(x) + \mu^* f_{a}^A(x)) v'(w^A(x)) - f^A(x)]^2 dx}} \propto T(x, \lambda^*, \mu^*). \]

First, let us characterize the solution to \((\text{Adj}_{\text{local}})\). Recall that the dual program is convex (even if the primal is not convex), because it is the pointwise minimum of affine functions. Therefore, the multipliers \( \lambda^* \) and \( \nu^* \) obtained in (7) and (8) are necessary and sufficient for an optimum in the dual problem.

We will now show that strong duality holds, and so \( t^* \) solves the primal problem given in \((\text{Adj}_{\text{local}})\). Towards this goal, let \( \Pi^* \) denote the optimal value of the primal problem. Weak duality implies that \( L(\lambda^*, \nu^*) \geq \Pi^* \). Moreover, it is straightforward to verify that \( t(\lambda^*, \nu^*) \) is feasible for \((\text{Adj}_{\text{local}})\), and \( \lambda^* \) and \( \nu^* \) is strictly positive if and only if the respective (primal) constraint binds. This implies that the objective of \((\text{Adj}_{\text{local}})\) evaluated at \( t(\lambda^*, \nu^*) \) is equal

\(^{18}\)If \( \nu = 0 \), then the integrand of (6) is linear in \( t \), and so \( L(\lambda, 0) = \infty \).
to $L(\lambda^*, \nu^*)$, and so it must be the case that $L(\lambda^*, \nu^*) \leq \Pi^*$. Therefore, we conclude that $L(\lambda^*, \nu^*) = \Pi^*$, which proves that strong duality holds, and the perturbation $t(\lambda^*, \nu^*)$ is optimal for $(\text{Adj}_{\text{local}})$.

To complete the proof, we show that $w^A$ is locally optimal if and only if $T(x, \lambda^*, \mu^*) = 0$ for all $x$. Clearly, $w^A$ is locally optimal if and only if the optimal perturbation is equal to zero for all $x$, which is true only if for some $\lambda' \geq 0$, we have $(\lambda' \hat{f} + \mu^* f^A_a) v'(w^A) = \hat{f}$ for every $x$. Suppose this is the case (for some $\lambda' \geq 0$). Integrating both sides with respect to $x$ and using that $\int f_a^A dx = 0$ implies that $\lambda' = \int f^A / v'(w^A) dx$. It is straightforward to verify that $t_{\lambda', \mu^*} \equiv 0$ solves (6), and $L(\lambda', \nu) = \nu$ for every $\nu$. 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.

Proof of Lemma 3.

We can write the agent’s maximized utility given contract $w$ as:

$$u(w) = \max_a \int v(w(x)) [g(x) + ah(x)] dx - c(a)$$

$$= \int v(w(x)) g(x) dx + \max_a \int v(w(x)) h(x) dx - c(a)$$

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

Next, define the function

$$V(I) = \max_a aI - c(a).$$

Since $\tilde{a}(I)$ is everywhere continuous, by the envelope theorem, $V$ is continuously differentiable, and we have

$$V'(I) = \tilde{a}(I).$$

By the fundamental theorem of calculus, for any $I$ and $\tilde{I}$,

$$V(I) - V(\tilde{I}) = \int_{\tilde{I}}^{I} \tilde{a}(i) di.$$
We therefore have
\[ u(w) - u(\tilde{w}) = \int v(w(x))g(x)\,dx + V(I(w)) - \int v(\tilde{w}(x))g(x)\,dx - V(I(\tilde{w})) \]
\[ = \int [v(w(x)) - v(\tilde{w}(x))]g(x)\,dx + \int_{I(\tilde{w})} \tilde{a}(i)\,di, \]

which establishes the first claim. The second claim is immediate. \qed

Proof of Proposition 3.

Recall that
\[ AB(w_A, w_B | P) = (f^A, f^B), \]
\[ AB(w_A, w_B | \tilde{P}) = (\tilde{f}^A, \tilde{f}^B). \]

Obviously, if the first statement is true, the second is also true.

Next, suppose that \( AB(w_A, w_B | P) = AB(w_A, w_B | \tilde{P}) \). It is immediate that \( f^i = \tilde{f}^i \), and so \( a_i = \tilde{a}_i \) for \( i \in \{A, B\} \). Solving the system of equations \( f^i(x) = g(x) + a^i h(x) \) and \( \tilde{f}^i(x) = \tilde{g}(x) + a^i \tilde{h}(x) \) for \( i \in \{A, B\} \) and all \( x \), it follows that \( g = \tilde{g} \) and \( h = \tilde{h} \). Moreover, Condition 1 implies that \( I(w^i | P) = I(w^i | \tilde{P}) \) for each \( i \in \{A, B\} \), and because \( \beta \) and \( \epsilon \) depend only on \( a(w) \) and \( I(w) \), we have \( \epsilon = \tilde{\epsilon} \) and \( \beta = \tilde{\beta} \), per the first statement. \qed