American Economic Review 2022, 112(1): 267–303
https://doi.org/10.1257/aer.20200732

A/B Contracts†

By George Georgiadis and Michael Powell*

This paper aims to improve the practical applicability of the classic theory of incentive contracts under moral hazard. We establish conditions under which the information provided by an A/B test of incentive contracts is sufficient for answering 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 (2018) 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. (*JEL* D82, D86, D91)

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 standard 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 toward answering it.

To introduce our main ideas and to illustrate two problems that the approach we develop 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 α, and your worker's average sales are a. Your

* Georgiadis: Kellogg School of Management, Northwestern University (email: g-georgiadis@kellogg.northwestern.edu); Powell: Kellogg School of Management, Northwestern University (email: mike-powell@kellogg.northwestern.edu). Stefano DellaVigna was the coeditor for this article. We are grateful to the editor, four anonymous referees, Iwan Barankay, Dan Barron, Simon Board, Hector Chade, Abdoulaye Ndiaye, Ben Golub, Dean Karlan, Eddie Lazear, Devin Pope, Nathan Seegert, Amanda Starc, Jeroen Swinkels, and Steve Tadelis, as well as to participants at several seminars and conferences for helpful comments. Finally, we thank Henrique Castro-Pires for excellent research assistance.

† Go to https://doi.org/10.1257/aer.20200732 to visit the article page for additional materials and author disclosure statements.

expected profits are therefore $\Pi = (m - \alpha)a$. If you were to marginally increase your piece rate, the effect on your profits would be

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

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

You know your gross profit margin, the current piece rate, and the current average sales. 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: What are their effort costs? If they work a little harder, what is going to happen to the distribution of their sales? These are questions you likely do not know the answer to, but importantly, they are questions you do not need to know the answer to if you are willing to run an experiment.

Suppose you decide to run an A/B test on your workforce. You randomly divide it into a treatment group 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 these data to estimate $da/d\alpha$, and you can use the expression above to determine whether you should marginally increase or decrease your piece rate.

This example illustrates 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).

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 output $f(\cdot|a)$, which we normalize so that the mean output is a. As in Holmström (1979), we assume that the agent's first-order condition characterizes 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.[1] 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 effort costs equal 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. The agent's behavioral response to a change in his marginal incentives, da/dI, is therefore independent of the adjustment to the contract that led to the change in marginal incentives. 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, da/dI, and how the adjustment affects his marginal incentives, dI.

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. The A/B test also provides the information required to estimate how *any other* marginal adjustment to the status quo contract affects the agent's marginal incentives, $\widetilde{dI}$, and therefore the agent's effort choice $\widetilde{da} = (da/dI)\widetilde{dI}$. A single A/B test, therefore, provides all the relevant information for predicting how the principal's expected profits will change in response to any marginal 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

[1] We write f_a to denote the derivative of $f(x|a)$ with respect to a, and we suppress the dependence on output x and effort a to simplify the notation.

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 a procedure for using 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 (2018) 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 performance 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 performance in each of the remaining four treatments and compare it to the actual average performance. A/B tests using piece-rate contracts predict the performance in the other piece-rate-contract treatments well: the mean absolute percentage error (APE) for such predictions is 0.66 percent. A/B tests using piece-rate contracts also predict the performance in bonus contracts well, and vice versa: The mean APE for such predictions is 2.28 percent. As a comparison, the mean absolute percentage performance differences across treatments is 6.40 percent. Moreover, our predictions for a given treatment are similar no matter which A/B test we use to make our predictions. Taken together, the correlation between our predictions and actual performance is 0.94.

Our second empirical exercise assesses the performance of the contract generated by our procedure. We use data from seven treatments to fit the parameters of the production environment using nonlinear least squares estimation. Given those parameters and an assumption about the principal's marginal revenue per unit of performance, we compute, as a benchmark, the optimal contract and the principal's corresponding expected profit. Then, we take data from each pair of treatments, and we use our procedure to construct the optimally adjusted contract. We define the *realized gains* of an adjustment to be the difference in profits between the adjusted and the status quo contract, and we define the *maximum gains available* to be the difference in profits between the optimal and the status quo contract. Averaging across all A/B tests, the realized gains are equal to approximately 68 percent of the maximum gains. Put differently, our results suggest that with a single A/B test, the principal can attain just over two-thirds of the profit gains that she could attain if she knew the entire production environment and put the optimal contract in place. We also demonstrate that this finding is robust to the principal's assumption about the agent's preferences for money.

Although our main results apply only to the canonical principal-agent framework of Holmström (1979), we show how our main insights extend to several enrichments of the framework. For example, we show how they extend to settings where the firm employs heterogeneous agents and to settings where the agent's effort is multidimensional.

Finally, we carry out both our empirical exercises in another experimental setting studied in DellaVigna and Pope (2021), where subjects perform a data-entry task under several different incentive schemes. First, we use each pair of incentive treatments to predict mean performance in each of the remaining treatments. Averaging across all pairs, the mean APE for such predictions is 5.14 percent, while the mean absolute percentage performance difference across treatments is 31.93 percent. In our second empirical exercise, we again construct a benchmark model and measure what fraction of the maximum gains available are realized by the test-optimal contract. Averaging across all A/B tests, the realized gains are approximately 75 percent of the maximum gains.

This paper contributes to both the theoretical and empirical literatures on principal-agent problems under moral hazard. Over the past four decades, theoretical work has extended the canonical principal-agent framework (Mirrlees 1976, Holmström 1979) to incorporate a host of additional real-world considerations: team production (Holmström 1982), dynamic incentives (Holmström and Milgrom 1987), limited liability (Innes 1990), multitask problems (Holmström and Milgrom 1991), behavioral agents (Bénabou and Tirole 2002, 2003, 2006), private information (Carroll 2015, Gottlieb and Moreira 2017, Chade and Swinkels 2019, Foarta and Sugaya 2021), and commitment problems (Laffont and Tirole 1988, MacLeod and Malcomson 1988). These papers characterize optimal contracts in their enriched settings and deliver deep insights into fundamental trade-offs. Their use as prescriptive theories has been limited, however, as the optimal contracts they prescribe in a given environment often depend in complicated and subtle ways on unobservable characteristics of that environment.

In order to take a step toward a prescriptive contract theory, we depart from much of the theoretical literature in two ways. First, we drop the strong assumption that the principal knows the production environment—the agent's effort-cost function and the joint distribution of effort and output. Second, instead of asking, "What is the best incentive contract?" we ask a narrower question, but one that is relevant in any ongoing organization: "What is the best way to improve upon an existing contract?" Our focus is on developing an understanding of what the principal needs to know—and equally important, what she might plausibly be able to know—to answer this question.[2] Carroll (2015) and Gottlieb and Moreira (2017) also assume the principal does not know the production environment. In contrast to these two papers, our focus is on how the principal can learn the relevant aspects of the environment, rather than their complementary approaches of describing optimal contracts when she cannot learn this information.

[2] Ortner and Chassang (2018) address a similar question in the context of designing policies to fight corruption: using a variational approach similar to ours, they show how a designer can use naturally occurring data to evaluate local policy changes.

Empirical work on incentive contracts has focused largely on testing key predictions of the theory. Several papers use quasi-experimental or experimental variation and show that higher-powered incentives cause workers to work harder, at least on the dimensions that are highly rewarded. These effects have been found in a variety of settings, ranging from windshield repairers (Lazear 2000), tree planters (Shearer 2004), and bicycle messengers (Fehr and Goette 2007) in high-income countries; to day laborers (Guiteras and Jack 2018), factory workers (Hong et al. 2018), and journalists (Balbuzanov, Gars, and Tjernström 2017) in low-income countries. Our results show that one could potentially use the data in each of these settings to improve upon the contracts being offered in that setting, subject to the caveat that any important discrepancies between the applied setting and the canonical moral-hazard setting we analyze would need to be accounted for in the analysis.[3]

This paper is related to the literature on sufficient statistics, which exploit envelope conditions from agents' optimization problems to characterize optimal policies in terms of simple elasticities and a small set of other model parameters; see Chetty (2009) for an overview and a unified framework, and Kleven (2020) for a generalization. In a seminal contribution, Harberger (1964) proposes a simple elasticity-based formula to measure the deadweight loss of a commodity tax. This approach has been used to study trade-offs in the design of monopoly pricing schemes (Wilson 1993), unemployment insurance (Baily 1978 and Chetty 2006), income-tax schedules (Feldstein 1999 and Saez 2001), welfare programs (Finkelstein and Notowidigdo 2019), and stimulus programs (Michaillat and Saez 2019).[4] Our paper extends the sufficient-statistics approach to analyze settings of pure moral hazard, where an agent's incentives depend on the entire contract he faces, and a change in his action affects the entire output distribution.

Finally, there are three papers that merit particular attention because they ask questions that are related to ours. Ke (2008) develops an approach for testing whether a contract is optimal using observational data on pay and performance under that contract. Our approach shows how experimental data can be used not only to test whether a given contract is optimal but how to improve upon it when it is not optimal. Prendergast (2015) shows how to bound the elasticity of workers' performance with respect to the output sensitivity of their pay by using information on their elasticity of taxable income. This information can inform whether a worker's pay should optimally be more sensitive to their performance, but it does not provide guidance for how best to adjust a worker's contract to achieve that goal. D'Haultfoeuille and Février (2020) use variation in contracts to estimate the losses associated with using linear contracts when workers are risk neutral. We show how such variation can be used to improve upon any suboptimal contract in more general pure-moral-hazard settings.

[3] For example, if learning by doing is an important source of productivity gains, the framework should be enriched to include an experience-dependent term to the agent's cost function.

[4] This approach can also be adapted to settings where envelope conditions are not applicable because, for example, agents are imperfect optimizers. See DellaVigna (2009) for a survey of evidence where individuals' behavior deviates systematically from the predictions of neoclassical optimization models; and Chetty, Looney, and Kroft (2009) for an application of the "sufficient statistics" approach to commodity taxation.

I. Model

We consider a standard contractual relationship between a principal and an agent as in Holmström (1979) but with a nonstandard 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 (hereafter pdf) $f(x \mid a)$, which we assume is twice continuously differentiable in a. Without loss of generality, we normalize a so that $a = E[x \mid a]$, and the agent's effort can be interpreted as his expected output.

If the agent is paid ω and chooses effort level a, he obtains utility $v(\omega) - c(a)$, where $v : \mathbb{R} \rightarrow \mathbb{R}$ and $c : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ 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, $f(\cdot \mid a(w))$. 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 the status quo contract. The set of contracts we allow the principal to choose from will depend on the exercise we carry out. In Section II, it will be the set of local adjustments to the status quo contract, and in Section III, it will be the full set of contracts.

Discussion: Our model aims to capture a setting where a firm employs a group of agents and has outcome data for the group of agents under two incentive contracts. This "many-agents" interpretation is fully consistent with our model as long as the agents are identical. In Section VA we establish conditions under which our results extend to settings with unobserved agent heterogeneity.

The assumption that the principal has outcome data from only two incentive contracts reflects the fact that experimenting with different incentive contracts can be very costly for firms. As we will show, outcome data from two incentive contracts provides all the information necessary to solve for the optimal local adjustment to the status quo contract in the classic moral-hazard setting. We also show that outcome data from additional test contracts may be useful if the agent's action is multidimensional (see Section VB) or for relaxing extrapolation conditions when solving for optimal non-local adjustments (see online Appendix A.B and A.C).

Throughout the paper, we take this outcome data as given and assume it has not been manipulated by agents either strategically to influence the principal's learning or nonstrategically, for example, if the unequal assignment to the test contract violates agents' fairness norms. In Section VI, we briefly discuss these and other issues that can arise when experimenting with agents' compensation and how they may be partially addressed through the appropriate design of high-level test-contract features that are outside the model.

Our focus on Pareto-improving contracts implicitly assumes that the agent's participation constraint binds under the status quo contract. An alternative rationale for this assumption is that when firms revise their performance-pay plans, workers are often suspicious about the firm's intentions, which can lead to opposition to the change or sabotage to its implementation; see, for example, Lazear (2000). Restricting attention to contracts that make workers at least as well off as the status quo contract may prevent these problems.

Finally, while our assumption that the principal knows the agent's utility function is restrictive, it is standard in both the contracting under moral hazard literature and the taxation literature; see, for example, Holmström (2017), the review of Chung, Kim, and Syam (2020); and Saez (2001). In each of our empirical exercises, we will assess the sensitivity of our results to the specific utility function we assume. We interpret this assumption as being consistent with the idea that managers can use information about the agent's decisions in other domains to learn about their risk preferences (see, for example, Einav et al. 2012 for evidence that individuals' risk preferences have a domain-general component). Moreover, the principal can also learn about the agent's utility function if she has outcome data from additional contracts.

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

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

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

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 $I(w,a) \equiv \int v(w(x)) f_a(x \mid a)\, dx$, where $f_a(x \mid a)$ is the derivative of $f(x \mid a)$ 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*

$$\mathcal{D}u(w,t) = \int t(x) v'(w(x)) f(x \mid a(w))\, dx$$

and his effort by

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

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

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 from the envelope theorem. The second part shows that the agent's behavioral response depends on how the adjustment affects his marginal incentives, $\mathcal{D}I(w,t)$, as well as on the local curvature of his problem. It also implies that $\mathcal{D}a(w,t)/\mathcal{D}I(w,t)$ 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.[5] Her expected profit under contract w is

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

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

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

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, holding expected output fixed.

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

[5] We assume that the principal is an expected profit maximizer, but it is straightforward to extend the results to any objective function that depends on the distribution of output.

the direction t that maximizes her profit differential subject to the constraint that it weakly improves the agent's utility. That is, she solves

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

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

In describing this problem, we temporarily assumed the principal knows the production environment. We now show she only needs to know certain local aspects of the production environment to solve (Adj_{local}). To do so, we will compare her problem across different production environments, and so it will be helpful to introduce the notation $(\text{Adj}_{local}{-}P)$ to refer to the principal's problem (Adj_{local}) when the production environment is P. Denote the agent's effort choice, the output density function, and 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 (Adj_{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* $(Adj_{local}{-}P)$ *if and only if it solves* $(Adj_{local}{-}\tilde{P})$.

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 pdf for w^A and f^B is the pdf 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 (Adj_{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:*

$$(i) \;\; f^A = \tilde{f}^A, f_a^A = \tilde{f}_a^A \text{ and } \mathcal{D}a(w^A, t\,|\,P) = \mathcal{D}a(w^A, t\,|\,\tilde{P}) \text{ for all } t.$$

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

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 (Adj_{local}). In particular, knowledge of f_a^A enables the principal to compute how the agent's marginal incentives change in response to adjusting the status quo contract in any direction, that is, $\mathcal{D}I(w^A, t)$ 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, we have for any t,

$$\mathcal{D}a(w^A, t) = \frac{\mathcal{D}a(w^A, w^B)}{\mathcal{D}I(w^A, w^B)} \mathcal{D}I(w^A, t).$$

Knowledge of $\mathcal{D}a(w^A, w^B)$, therefore, allows the principal to evaluate the agent's effort differential as the status quo contract is adjusted in any direction t and ultimately solve (Adj_{local}).

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

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

Proposition 2 characterizes the optimal local adjustment.

PROPOSITION 2: *Let w^A be the status quo contract. There exist $\lambda^*, \mu^* \geq 0$ such that $t^*(x) \propto T(x, \lambda^*, \mu^*)$ solves (Adj_{local}). 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 and 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 toward 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 λ^*, μ^*, the exact expressions for which are given in the proof of Proposition 2 in the online Appendix.

The second part of this proposition echoes the optimality conditions of Holmström (1979) and serves as a consistency check. When the status quo contract is already optimal, the coefficients λ^* and μ^* coincide with those in Holmström (1979). The primary contribution of Proposition 2 is to show how λ^* and μ^* change as we consider status quo contracts that are not locally optimal. In particular, μ^*, the weight that is optimally put on how marginal incentives are adjusted, is higher when the principal's expected gains from a higher effort level are higher and when the agent's response to an increase in marginal incentives is higher. The weight that is put on the risk-allocation component, λ^*, is smaller when μ^* is higher.

III. Non-local Adjustments

The analysis in Section II illustrates how local information suffices for characterizing optimal local adjustments. This section provides a method for extrapolating to assess non-local adjustments. It shows in particular how to use non-local 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 $f_a(x \mid a^A)$ and $\mathcal{D}a(w^A, t)$, 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 $f(x \mid a)$ for all a and $a(w)$ for all w. This section provides a pair of conditions under which this information can be extrapolated from a single A/B test. Throughout, we focus on a specific set of extrapolation conditions, which are the ones we use in our empirical exercises in Section IV. At the end of this section, we discuss more general conditions that suffice for extrapolation from 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 common in the moral-hazard literature because it guarantees the first-order approach is valid. It also implies several further properties that are useful for our exercise. First, it ensures that knowledge of the pdf, $f(\cdot \mid a)$, at two effort levels, say a^A and a^B, is sufficient to estimate the pdf corresponding to any other effort level. Second, it implies that this information also suffices to compute $f_a(x \mid a) \equiv h(x)$ 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. Additionally, it ensures that $f(x \mid a)$ does not have a moving support, which could lead to optimal contracts that depend critically on this property. One limitation of imposing this extrapolation condition is that the output distribution can be computed only for efforts such that $g(x) + ah(x) \geq 0$ for all x.

We will now revisit the agent's problem, under the assumption that Condition 1 is satisfied. Given a contract w, he solves

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

The agent's optimal effort level, given marginal incentives I, which we denote by $\tilde{a}(I)$, satisfies the implicit equation $c'(\tilde{a}(I)) = I$. The following lemma parallels Lemma 1 and characterizes the agent's utility and effort when contract w is replaced with contract $\widetilde{w}$.

LEMMA 3: *Suppose Condition 1 is satisfied, and the contract w is replaced with $\tilde{w}$. Then the agent's utility satisfies*

$$u(\widetilde{w}) = u(w) + \int \left[v(\widetilde{w}(x)) - v(w(x))\right] g(x)\,dx + \int_{I(w)}^{I(\widetilde{w})} \tilde{a}(i)\,di$$

and his effort satisfies

$$a(\widetilde{w}) = a(w) + \int_{I(w)}^{I(\widetilde{w})} \frac{d\tilde{a}(i)}{di}\,di,$$

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

$$\ln a(w) = \beta + \varepsilon \ln I(w). \tag{3}$$

An A/B test provides the information required to determine β and ε. 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 equals 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 β can be constructed using this information as well: $\beta = \ln a^A - \varepsilon \ln I^A$. This condition ensures, therefore, that the agent's effort choice can be extrapolated given information on a single behavioral elasticity, which is consistent with the standard approach taken in the sufficient statistics literature for optimal taxation; see, for example, Brewer, Saez, and Shephard (2010).[6]

Let us now define the principal's profit when she offers contract $\widetilde{w}$,

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

The principal's problem given the status quo contract w^A is therefore

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

[6] We implicitly assume that both the status quo and test contracts generate strictly positive marginal incentives, precluding, for instance a constant-wage contract. Moreover, Condition 2 implies that $c'(0) = 0$ and therefore such a contract would motivate zero effort, which is at odds with evidence from many settings, including the one we will study in the next section. To accommodate positive effort choices under constant-wage contracts, we can add a parameter to the agent's cost function that captures incentives that are external to the model, such as those arising from intrinsic motivation or long-term career incentives. These external incentives can be identified with outcome data from an additional test contract. See online Appendix B for details.

In practice, the program (Adj) is solved using the Grossman and Hart (1983) two-step approach. In the first step, we fix a target effort level a and solve for the cost-minimizing contract that satisfies $a(\widetilde{w}) = a$ and $u(\widetilde{w}) \geq u(w^A)$. In the second step, we choose the optimal target effort level. The first-stage problem can be transformed into a convex program by transforming the principal's choice from the function $\widetilde{w}$ to the function $V = v(\widetilde{w})$. 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.

Under Conditions 1 and 2, the principal can learn all the relevant parameters of the production environment with an A/B test, allowing her to solve (Adj). We now formally state this result, which is the sufficient-statistic analogue of Proposition 1 for non-local adjustments. Similar to Section II, we will write (Adj$-P$) to refer to the principal's problem (Adj) when the production environment is P.

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$-\tilde{P}$).

This proposition shows that when Conditions 1 and 2 hold, an A/B test provides the necessary information to solve the principal's problem (Adj). We also note that this sufficient-statistic result continues to hold if the problem (Adj) is augmented with additional constraints that depend only on the contract $\widetilde{w}$, such as limited-liability or monotonicity constraints. We use this observation in our second empirical exercise.

We conclude with a brief discussion of the extrapolation conditions. As Proposition 3 shows, these conditions are sufficient to ensure that the information from a single A/B test allows the principal to calculate an optimal contract. Condition 1 amounts to linearly extrapolating, for every x, the points $(a^A, f^A(x))$ and $(a^B, f^B(x))$ to compute $f(x\,|\,a)$ for a range of a's.[7] Condition 2 amounts to an isoelastic extrapolation of the pair of effort levels and marginal incentives in an A/B test. Analogous results hold under the following more general extrapolation condition that makes use of data from two contracts.

CONDITION 2': *The agent's marginal cost function* $c'(a;\theta_1,\theta_2)$ *is such that there is a unique pair of parameters* θ_1 *and* θ_2 *satisfying* $c'(a^A;\theta_1,\theta_2) = I(w^A,a^A)$ *and* $c'(a^B;\theta_1,\theta_2) = I(w^B,a^B)$ *for any pair of contracts* w^A *and* w^B.

[7] In principle, this condition can be replaced by any extrapolation method that makes use of the data from only two contracts. For example, one might instead assume that $f(x|a) = \widetilde{g}(x) + a\widetilde{h}(x) + k(a)i(x)$ for some known functions $k(a)$ and $i(x)$. Using data from two contracts, it is possible to recover $\widetilde{g}(x)$ and $\widetilde{h}(x)$ and analogous results hold.

Condition 2 is a special case of Condition 2' with $c'(a;\theta_1,\theta_2) = e^{-\theta_2/\theta_1} a^{1/\theta_1}$. As an example, one might instead assume a cost function of the form $c'(a;\theta_1,\theta_2) = \theta_1 e^{\theta_2 a}$, which satisfies Condition 2' but not Condition 2 (DellaVigna and Pope 2018).

IV. An Empirical Exploration

We will now assess the quantitative implications of our model. To do so, we use data from DellaVigna and Pope's 2018 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 countdown 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 treatment 2, they were paid a constant amount for every thousand points, and in treatments 3 to 5, they were paid a constant amount for every hundred points. 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*. For consistency with our model, we treat x as a continuous variable. Therefore, the implied incentive contracts for these treatments are an approximation. 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 $w^6(x) = 100 + 40\,\mathbb{I}_{\{x\geq 2000\}}$, where $\mathbb{I}_{\{x\geq 2000\}}$ 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 data from two treatments to predict the performance in the remaining treatments. The second exercise assesses the performance of the optimal adjustment generated by our procedure relative to a benchmark that we construct from the data using the treatments in Table 1.

A. *Predicting Out-of-Sample Experimental Results*

Our results in Section III show how to use outcome data from two contracts to predict agents' effort under an arbitrary contract. We will assess the accuracy and

Table 1—Experimental Treatments from DellaVigna and Pope (2018)

	Contract	Average number of points	Standard deviation	Number of subjects
No incentives	$w^1(x) = 100$	1,521	726	540
Piece rate	$w^2(x) = 100 + 0.001x$	1,883	664	538
	$w^3(x) = 100 + 0.01x$	2,029	649	558
	$w^4(x) = 100 + 0.04x$	2,132	626	562
	$w^5(x) = 100 + 0.10x$	2,175	578	566
Bonus	$w^6(x) = 100 + 40\mathbb{I}_{\{x \geq 2{,}000\}}$	2,136	576	545
	$w^7(x) = 100 + 80\mathbb{I}_{\{x \geq 2{,}000\}}$	2,188	530	532

Notes: This table describes seven experimental treatments from DellaVigna and Pope (2018) 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.

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.

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 ε and β.[8] We will also assume that the agent has constant-relative-risk-aversion (CRRA) 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.[9]

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.[10] The outcome data for treatment j are a cumulative distribution function F^j. For each treatment j,

[8] While it is likely that subjects differ in various dimensions such as their ability or willingness to perform repetitive tasks, we are unable to estimate any subject-specific heterogeneity, because each subject participated only once. As such, we treat subjects as being homogeneous, and we use our baseline model to make our predictions. Section A provides conditions under which doing so is without loss of generality.

[9] We assume narrow framing, that is, that subjects do not integrate the experimental earnings with any other part of the their portfolio. Otherwise, even if they are risk averse, their marginal utility would, in effect, be constant over such small payoffs.

[10] 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

we use a kernel density estimator to construct the pdf $\hat{f}^j$.[11] Then, for each pair (A, B), we use these pdfs to construct the function

$$\hat{h}^{AB}(x) = \frac{\hat{f}^A(x) - \hat{f}^B(x)}{a^A - a^B}. \tag{4}$$

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

$$\hat{I}_C^{AB} = \int v\big(w^C(x)\big)\,\hat{h}^{AB}(x)\,dx.$$

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 we suppose to be the status quo and test contracts.[12] Finally, our prediction for average points accumulated in treatment C is $\ln \hat{a}_C^{AB} = \hat{\beta}^{AB} + \hat{\varepsilon}^{AB} \ln \hat{I}_C^{AB}$.

We focus first on what we refer to as **homogeneous A/B tests**, A/B tests in which treatments A and B are in the same class; that is, 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*,

(*i*) *predicted out-of-sample performance is highly correlated with actual performance*,

(*ii*) *predictions are close to actual performance for both within-class and across-class predictions*, *and*

(*iii*) *predictions for a given treatment are similar no matter which pair of contracts is used to construct the prediction.*

---

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 IVB and online Appendix B.

[11] 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 ten-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.

[12] That is because these objects are symmetric in (A, B): $\hat{g}^{AB} = \hat{g}^{BA}$, $\hat{\varepsilon}^{AB} = \hat{\varepsilon}^{BA}$, and $\hat{\beta}^{AB} = \hat{\beta}^{BA}$.

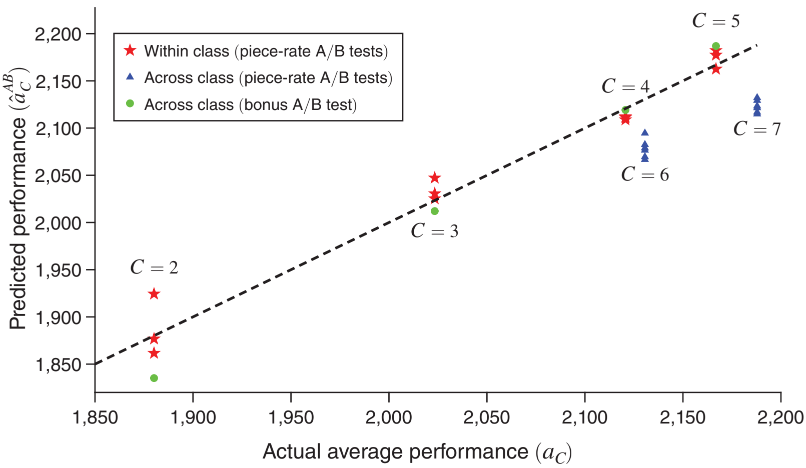

Figure 1

Notes: 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 performance, a_C, for treatments $C \in \{2, 3, \ldots, 7\}$, while the vertical axis plots predicted performance, $\hat{a}_C^{AB}$. The red stars represent predictions of piece-rate treatments using A/B tests from other piece-rate treatments. The blue triangles represent predictions of bonus treatments using A/B tests from piece-rate treatments. The green circles represent predictions of piece-rate treatments using the A/B test from the bonus treatments.

For all homogeneous A/B tests, Figure 1 plots our predictions against the actual average performance for each treatment. The horizontal axis depicts the actual average performance, a_C, for treatments $C \in \{2, \ldots, 7\}$, 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, which is Result 1(i).

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

$$APE\left(\hat{a}_C^{AB}\right) = \left|\frac{\hat{a}_C^{AB} - a_C}{a_C}\right|.$$

The mean 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 mean 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 mean APE is 0.99 percent. The predictions are slightly worse when A and B are piece-rate treatments, and C is a bonus treatment. There, the mean APE is 2.71 percent, and as Figure 1 shows, they systematically underestimate performance. We discuss our interpretation of this pattern at the end of this section. Notice, however, that all predictions are close to the 45-degree line, depicted by the dashed line, illustrating Result 1(ii).

Finally, Figure 1 also shows that the estimates of each treatment's performance are tightly clustered, illustrating Result 1(iii). To quantify this result, we can compute,

Table 2—Out-of-Sample Effort Predictions

Coefficient of RRA (ρ)	0 (1)	0.1 (2)	0.2 (3)	0.3 (4)	0.4 (5)	0.5 (6)	1[a] (7)
Panel A. Homogeneous A/B tests							
$\mathrm{Corr}(\hat{a}_C^{AB}, a_C)$	0.92	0.93	0.94	0.94	0.95	0.96	0.97
Mean APE (percent)	1.76	1.69	1.62	1.59	1.56	1.54	1.64
Within class	0.84	0.76	0.67	0.66	0.67	0.67	1.06
Across class: piece-rate predictions	1.01	0.99	0.97	0.99	1.05	1.11	2.15
Across class: bonus predictions	2.93	2.86	2.79	2.71	2.63	2.55	2.04
Worst-case APE (percent)	3.65	3.56	3.45	3.34	3.21	3.08	4.30
Within class	3.43	3.10	2.74	2.35	1.92	1.45	2.56
Across class: piece-rate predictions	1.76	1.90	2.14	2.39	2.64	2.90	4.30
Across class: bonus predictions	3.65	3.56	3.45	3.34	3.21	3.08	3.03
Average CV of estimates (percent)	0.82	0.78	0.74	0.70	0.68	0.68	0.83
Panel B. Hybrid A/B tests							
$\mathrm{Corr}(\hat{a}_C^{AB}, a_C)$	0.86	0.86	0.85	0.84	0.84	0.83	0.78
Mean APE (percent)	2.19	2.18	2.17	2.16	2.15	2.14	2.18
Worst-case APE (percent)	10.60	10.63	10.66	10.70	11.07	11.40	12.69
Average CV of estimates (percent)	2.03	2.03	2.04	2.05	2.05	2.06	2.09

Notes: This table reports summary statistics for predicted performance under different assumptions for the agent's coefficient of RRA. Column 4 represents our baseline assumption that $\rho = 0.3$, and the remaining columns vary ρ. Panel A reports, for homogeneous A/B tests, the correlation between predicted and actual performance, the mean and worst-case absolute percentage error (APE), and the coefficient of variation (CV) of the estimates. Panel B reports these quantifies for the hybrid A/B tests.

[a] Unit coefficient of RRA corresponds to the logarithmic utility function; i.e., $v(\omega) = \ln \omega$.

for each treatment C, the coefficient of variation of the predictions $\hat{a}_C^{AB}$. The average coefficient of variation across the six treatments is 0.7 percent and ranges between 0.21 percent for treatment 3 and 2 percent for treatment 2.

These results are summarized in Table 2, panel A, column 4. This panel also shows two additional results. First, the worst-case APE, defined as $\max APE(\hat{a}_C^{AB})$, is also small. This is true for both 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 (hereafter RRA). The prediction accuracy is also similar if the agent's utility is assumed to belong to a different class of functions.[13] Figures 8 and 9 in online Appendix A.A compare the predicted and the empirical output distribution for each treatment and every homogeneous A/B test.

Result 1 and Table 2, panel A focus on homogeneous A/B tests. We now discuss our predictions using hybrid A/B tests, which are summarized in Table 2, panel B. Across all predictions involving hybrid A/B tests, the correlation between $\hat{a}_C^{AB}$ and a_C is 0.84, and the mean 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 hybrid (A, B) pairs that perform particularly poorly are $(4, 6)$, $(5, 6)$, and $(5, 7)$.

[13] If, for example, $v(\omega) = 1000\omega - b\omega^2$, and we vary b from zero to one (thus ensuring that marginal utility is always nonnegative), the mean APE varies between 1.44 percent and 1.76 percent for homogeneous A/B tests, and between 2.04 percent and 2.19 percent for hybrid A/B tests.

To see why, let us focus on the $(5,7)$ pair—the lessons are similar when we look at $(4,6)$ and $(5,6)$. The output distributions under these two treatments have distinctly different patterns, as illustrated in the left panel of Figure 11. In particular, for treatment 5, which is a piece-rate treatment, performance is roughly symmetrically distributed around the average. For the bonus treatment 7, however, performance spikes just over $x = 2000$, the threshold for receiving the bonus. This is because in contrast to our model where effort is chosen once and for all, in the experiment, subjects can adjust their effort over time.[14] The estimated function $\hat{h}^{AB}$ magnifies these differences, because the average performance in these two treatments is quite similar, with $a_5 = 2175$ and $a_7 = 2187$, and this difference appears in the denominator of (4). For, say, the $(2,7)$ pair, we see similarly distinct patterns. Since the average performance in treatment 2, $a_2 = 1883$, is significantly lower than in treatment 7, however, our out-of-sample predictions are less influenced by these patterns.

The reason why A/B tests comprising piece-rate treatments underpredict the performance of the bonus treatments is related. The function $\hat{h}^{AB}$ constructed using bonus treatments tends to take large positive values for x just over 2000, which is the threshold for receiving the bonus and small or negative values for other values of x. As a result, the implied marginal incentives generated by a contract that 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 for x values just over 2000. Predictions of bonus-treatment performance constructed using output data from piece-rate treatments systematically underpredict the marginal incentives, and hence the effort, generated by bonus contracts, although only by about two percent.

B. *Performance of Optimal Adjustments*

For our second exercise, we will assess the empirical performance of our solution to the principal's problem (Adj). To do so, we must first develop a benchmark to compare it against. For this, we will again use DellaVigna and Pope's (2018) data and will proceed in two steps. First, we will build a benchmark model using the data from several of the treatments. Then for each treatment C, we will compute the benchmark-optimal contract 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 test-optimal contract that solves (Adj) and gives the agent at least as much expected utility as w^C. We will then compare its performance to that of the corresponding benchmark-optimal contract. In light of our results in Section IVA, we focus on (A, B) pairs from the same class in this section, and report results for hybrid A/B tests in online Appendix A.A.

The Benchmark Model and Optimal Adjustments.—We now describe how we construct our benchmark model. Throughout, we will use tildes to denote compo-

[14] Online Appendix A.C considers an extension in which subjects are allowed to choose the entire output distribution.

nents of the benchmark model. First, we construct the benchmark pdf $\tilde{f}(x|a)$ for all $x \in [0, 3500]$ and for all a within a particular interval, which we will describe below. Next, we construct the parameters of the agent's cost function. As in the previous section, we will assume that the agent has 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.

Benchmark Output Distribution: To construct the benchmark pdf $\tilde{f}(x \mid a)$, we proceed in two steps. First, we use outcome data for treatments 1 to 5—the *no-incentives treatment* and the *piece-rate treatments*. We discuss this choice in footnote 17. These outcome data are a set of cumulative distribution functions $F(x \mid a^C)$, 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 pdf $\hat{f}(x \mid a^C)$ for each treatment $C \in \{1, \ldots, 5\}$.[15] We assume that $\tilde{f}(x|a) = \hat{f}(x|a)$ for all $a \in \{a^1, \ldots, a^5\}$, and for each x, we use a spline interpolation to construct $\tilde{f}(x|a)$ for other values of a between a^1 and an upper bound, $\bar{a}$. The spline interpolation is not guaranteed to satisfy $\tilde{f}(x \mid a) \geq 0$ 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 $\tilde{f}(x \mid a) \geq 0$ for all $a \in [a^1, \bar{a}]$ for all x. Finally, given the benchmark pdf $\tilde{f}(x \mid a)$, we approximate its derivative as $\tilde{f}_a(x \mid a) = \tilde{f}(x \mid a+1) - \tilde{f}(x \mid a)$.

Agent's Benchmark 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 $\tilde{c}'(a) = e^{-\tilde{\beta}/\tilde{\varepsilon}} a^{1/\tilde{\varepsilon}} - \tilde{I}_0$ 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 benchmark cost function therefore requires fitting three parameters to the data: $\tilde{\varepsilon}, \tilde{\beta}$, and $\tilde{I}_0$. Table 3 reports the fitted values for these parameters using nonlinear least squares estimation.[16]

Benchmark-Optimal Contract: We then solve for the principal's benchmark-optimal contract. Recall that the benchmark-optimal contract depends on what the status quo contract is because it determines the utility that the principal

[15] Again, we use the triweight kernel estimator with the bandwidth determined by the Silverman rule of thumb, and have excluded observations with $x > 3500$.

[16] For each treatment C, we compute $\tilde{I}^C = \int v(w_i(x)) \tilde{f}_a(x|a^C)\, dx$ and minimize $\sum_{C=1}^{7} \left[\log(a_i) - \beta - \epsilon \log(\tilde{I}^C + I_0)\right]^2$ to obtain $\tilde{\varepsilon}, \tilde{\beta}$, and $\tilde{I}_0$. Constructing $\tilde{f}(x \mid a)$ using outcome data from only treatments 1 to 5 leads to a lower value for the minimized objective than constructing it with data from any other subset of the seven treatments.

Table 3—Fitted Parameters for the Benchmark Model

$\tilde{\varepsilon}$	$\tilde{\beta}$	$\tilde{I}_0$
0.0322	7.8184	6.528×10^{-7}

Notes: This table displays the fitted parameters for the benchmark model. They are computed using a nonlinear least squares estimation procedure.

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 **benchmark-optimal contract**, $w^*(w^C)$, by solving the following two-step problem.

First, for each integer $a \in [a^1, \bar{a}]$, we find the cost-minimizing contract that solves

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

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

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

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

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

and two additional constraints. First, we impose the constraint that $w(x) \geq 100$ for all x 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.[17,18,19]

For the second step, we do a line search to solve for the principal's optimal choice of a:

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

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

[17] Since we are not imposing Condition 1 in the benchmark model, the first-order approach is not always valid. We therefore impose a global incentive compatibility constraint, requiring that the target effort level gives the agent a larger expected utility than any other (integer) effort level.

[18] We impose the monotonicity constraint for two reasons. First, without it, the benchmark-optimal effort is always equal to the upper bound, $\bar{a}$, which implies that any test-optimal contract will mechanically implement an effort that is weakly smaller than is benchmark-optimal, limiting what we can learn from this exercise. Second, nonmonotonic contracts can motivate gaming and other undesirable behaviors (see, for example, Innes 1990 and Oyer 2000), and presumably for this reason, are hardly ever used in practice.

[19] We solve this problem with the CVX software for Matlab (Grant and Boyd 2013) after using the transformation $V(x) \equiv \tilde{v}(w(x))$ to convert it into a convex optimization program.

TABLE 4—PERFORMANCE OF OPTIMAL ADJUSTMENTS AND SENSITIVITY ANALYSIS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Model coefficient of RRA ($\widetilde{\rho}$)	0.3	0.3	0.3	0.1	0.5	0.3	0.3
Test coefficient of RRA ($\widehat{\rho}$)	0.3	0.3	0.3	0.3	0.3	0.1	0.5
Profit margin ($\widetilde{m}$)	0.2	0.15	0.25	0.2	0.2	0.2	0.2
Average gains ($)	7.17	5.21	9.26	7.25	7.15	6.85	7.31
Maximum gains ($)	10.55	7.62	13.52	10.74	10.59	10.55	10.55
Gains ratio (percent)	68.01	68.44	68.51	67.47	67.53	64.98	69.32
Average effort deviation	−6.74	−7.45	−6.31	−7.76	−6.55	−8.43	−6.57
Average overpayment ($)	1.82	1.32	2.25	1.78	2.02	2.13	1.64

Notes: This table reports for different values of the parameters $\tilde{\rho}$, $\hat{\rho}$, and $\widetilde{m}$, the average and maximum gains, the gains ratio, the average effort deviation, and the average overpayment, averaged across $C \in \{2, \ldots, 7\}$. Column 1 represents our baseline parameters. In columns 2 and 3 we vary the profit margin, $\widetilde{m}$. In columns 4 and 5 we vary the coefficient of RRA used in the benchmark model, $\widetilde{\rho}$. Finally, in columns 6 and 7 we vary the coefficient of RRA that the principal assumed to solve for the test-optimal contract given an A/B test, $\hat{\rho}$.

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

Test-Optimal Contracts: We then solve for the principal's test-optimal 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 a pdf 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 a pdf $\hat{f}^{AB}$ that satisfies $\hat{f}^{AB}(x|a) = \hat{g}^{AB}(x) + a\hat{h}^{AB}(x)$ for all x and for all $a \in [\underline{a}^{AB}, \bar{a}^{AB}]$, where $\underline{a}^{AB}$ and $\bar{a}^{AB}$ are chosen so that $\hat{f}^{AB}(x|a) \geq 0$ for all x and for all a in that interval. The cost-function parameters $\hat{\varepsilon}^{AB}$ and $\hat{\beta}^{AB}$ are constructed as in the previous section, assuming the agent's cost function satisfies $\hat{c}^{AB\prime}(x) = e^{-\hat{\beta}^{AB}/\hat{\varepsilon}^{AB}} a^{1/\hat{\varepsilon}^{AB}}$. We again assume that the agent has CRRA preferences over money $\hat{v}'(\omega) = \omega^{-\hat{\rho}}$ with $\hat{\rho} = 0.3$.

We then solve for the principal's test-optimal contract by solving the following two-step problem. First, for each integer $a \in [\underline{a}^{AB}, \bar{a}^{AB}]$, we find the cost-minimizing contract that solves

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

subject to the agent's first-order condition for effort

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

the constraint that the principal predicts the agent will be at least as well off as under the status quo contract

$$\int \hat{v}(w(x)) \hat{f}^{AB}(x\,|\,a)\, dx - \hat{c}^{AB}(a) \geq \int \hat{v}(w^C(x)) \hat{f}^{AB}(x|\hat{a}_C^{AB})\, dx - \hat{c}^{AB}(\hat{a}_C^{AB}),$$

as well as the two additional constraints we imposed when we solved for the benchmark-optimal contract: $w(x) \geq 100$ for all x and $w(x)$ is weakly increasing in x.[20,21]

For the second step, we do a line search to solve for the principal's optimal choice of a:

$$\max_{a \in [a^1, \bar{a}^{AB}]} \tilde{m}a - \hat{K}^{AB}(a; w^C).$$

Solving this problem gives us the test-optimal contract $w^{AB}(w^C)$. We then use the benchmark model to evaluate the agent's effort choice and the principal's expected profits under this contract. We refer to the resulting effort level, $a^{AB}(w^C)$, as the test-optimal effort level and the resulting profits, $\pi^{AB}(w^C)$, as the principal's test-optimal profits.

Performance.—We will now discuss the performance of test-optimal contracts. To do so, we first have to define what it means for test-optimal contracts to perform well. In particular, we will start with a status quo contract and compare how much the principal's expected profits increase when she puts in place the test-optimal contract to how much they increase when she puts in place the benchmark-optimal 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 available gains for treatment** C to be the difference between the benchmark-optimal profits and the status quo profits, that is,

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

where $\pi(w^C)$ is the expected profits in the benchmark model under status quo contract w^C. Second, we will define the **average realized gains for treatment** C to be the average difference between the test-optimal profits under status quo contract w^C across all homogeneous A/B tests and the status quo profits, that is,

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

where $Hom \equiv \{(A,B) | (A,B) \text{ is a homogeneous pair}\}$ and $|Hom| = 7$ because there are seven homogeneous A/B tests. Finally, we will define the **gains ratio** to be the sum over C of the average realized gains for treatment C divided by the sum over

[20] An implication of Condition 2 is that the first-order approach is valid. It is therefore without loss of generality to replace the agent's incentive compatibility constraint with the corresponding first-order condition.

[21] In principle, the agent's effort under the contract w^C should appear in the right-hand side of the agent's participation constraint. Of course, this quantity is not directly observed by the principal unless the A/B test contains treatment C. Therefore, we use the predicted effort under treatment C given the A/B test at hand, $\hat{a}_C^{AB}$, as described in Section IVA. The predicted effort is equal to the true effort if $C \in \{A, B\}$.

C of the maximum available gains for treatment C. The following result summarizes our main findings for the performance of test-optimal contracts.

RESULT 2: *For homogeneous A/B tests,*

(*i*) *the average gains ratio across treatments is about 68 percent,*

(*ii*) *approximately two-fifths of the gap between realized and maximum gains is due to the test-optimal contract implementing suboptimal effort, with the remainder attributable to implementing this effort at too high a cost.*

The first part of Result 2, which is illustrated in Figure 2, shows that test-optimal contracts perform well. The quantity $(1/6)\sum_{C=2}^{7}\text{AvgGains}^C$ is \$7.14, about 68 percent of the quantity $(1/6)\sum_{C=2}^{7}\text{MaxGains}^C$, which is \$10.55. In other words, the information from a single A/B test allows the principal to realize about 68 percent of the profit gains that she could achieve if she knew the production environment and could therefore compute the benchmark-optimal contract. The gap between the average realized gains and maximum available gains, $\text{MaxGains}^C - \text{AvgGains}^C$, is about \$3, and it exhibits little variation across treatments C. This is illustrated by the ordinary least squares fitted line in Figure 2, which has a slope and intercept close to 1 and -3, respectively, and is close to each of the points.

The second part of Result 2 sheds light on the sources of this gap. First, test-optimal effort levels tend to be close to but slightly smaller than benchmark-optimal effort levels. Second, test-optimal contracts tend not to be the cost-minimizing contracts for the effort levels they induce. Figure 3 below compares the test-optimal effort levels to the benchmark-optimal ones. On the horizontal axis, it plots the **benchmark-optimal effort change**, $a^*(w^C) - a^C$, for each treatment. On the vertical axis, it plots the average **test-optimal effort change** across all homogeneous A/B tests, that is, $(1/|Hom|)\sum_{A,B\in Hom}a^{AB}(w^C) - a^C$, for each treatment.

This figure illustrates several points. First, the benchmark-optimal effort change varies widely across treatments. For treatments 5 and 7, the benchmark-optimal effort change is negative, and for treatment 2, it is almost 200 points. Second, the average test-optimal effort change is close to the benchmark-optimal effort change; that is, $(1/|Hom|)\sum_{A,B\in Hom}a^{AB}(w^C) - a^C$ is close to the 45-degree line for each C. Averaging across all six treatments, the **average effort deviation**, which we define to be the difference between the benchmark-optimal effort change and the test-optimal effort change is -6.74: On average, test-optimal effort levels are 6.74 below the benchmark-optimal effort levels. Given that each unit of effort yields $\tilde{m} = 0.2$ dollars in profits for the principal, on average, the principal is losing about \$1.35 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 test-optimal contract $w^{AB}(w^C)$ induces effort level $a^{AB}(w^C)$ and therefore costs the principal

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

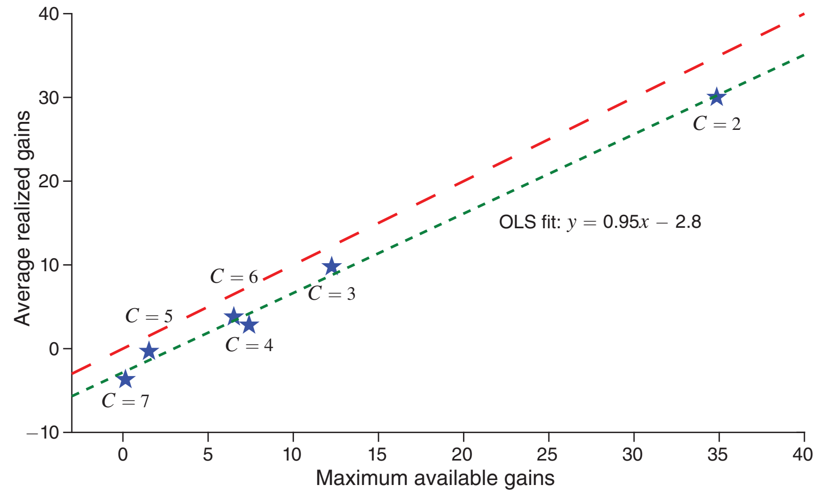

FIGURE 2

Notes: This figure compares the average realized gains to the maximum available gains for each treatment C. By construction, the average realized gains lie below the 45-degree line, which is depicted by the dashed red line. The green dotted line represents the ordinary least squares (OLS) regression line through the points $\left(\text{MaxGains}^C, \text{AvgGains}^C\right)$.

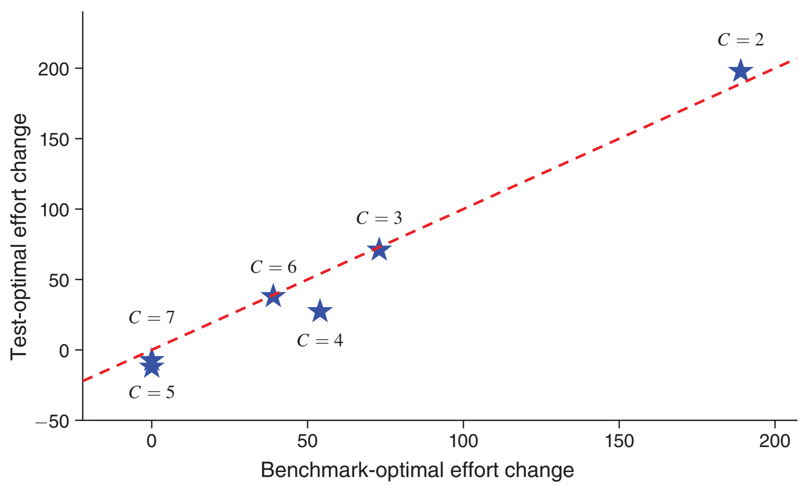

FIGURE 3

Notes: This figure compares, for each treatment C, the benchmark-optimal effort change and the test-optimal effort change. Each star represents the point with x-coordinate $a^*(w^C) - a^C$ and y-coordinate $(1/|Hom|)\sum_{A,B\in Hom} a^{AB}(w^C) - a^C$, for some treatment C. The dashed red line is the 45-degree line.

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

$- K\left(a^{AB}\left(w^C\right); w^C\right)$. Across the six treatments, the average overpayment is about \$1.82.

Table 4 reports these summary statistics for different values of the coefficient of RRA we used in the benchmark model, $\widetilde{\rho}$, the coefficient of RRA that the principal assumed to solve for the test-optimal contract given an A/B test, $\widehat{\rho}$, and the principal's profit margin, $\widetilde{m}$. Both average and maximum gains increase with $\widetilde{m}$, but the gains ratio exhibits little variation. Moreover, all summary statistics are relatively insensitive to the values of $\widetilde{\rho}$ and $\widehat{\rho}$.[22]

Online Appendix A.A reports additional results on the optimal adjustments. In particular, it presents disaggregated data for the optimal adjustment from each homogeneous A/B test, it illustrates some benchmark-optimal and test-optimal contracts, it reports summary statistics for the performance of optimal adjustments using hybrid A/B tests, and it carries out a robustness exercise when stakes are magnified.

V. Beyond the Classic Model

Our main analysis was carried out in the context of the classic setting of Holmström (1979). We showed in our second empirical exercise how the framework we developed could accommodate several additional considerations that were relevant to the experimental setting we analyzed, such as external incentives, limited-liability constraints, and monotonicity constraints.

In this section, we show how to extend our analysis in two additional directions. First, we show how to incorporate unobserved worker heterogeneity. We provide conditions under which the aggregate data contained in an A/B test suffices to predict workers' heterogeneous behavioral responses, and we quantitatively explore the discrepancies that arise when these conditions are not satisfied. Second, we consider settings in which the agent's effort and output are multidimensional. Effort substitution patterns become important for optimal adjustments, and we show that they can be identified with additional test contracts.

A. *Heterogeneous Workers*

Up to this point, we assumed that the principal faces a mass of identical agents, and we showed how she can use *aggregate* data on their performance to improve upon a status quo contract. In this section, we continue to assume that the principal has access to aggregate data generated by agents under a pair of contracts, but we now assume that these agents are heterogeneous. In particular, suppose there is a finite set of types, Φ, and agents with different types have different effort-cost functions but are otherwise identical.

There are two challenges that arise in general when using aggregate data from an A/B test to predict how a mass of heterogeneous agents will respond to a change in the contract. First, a given contract may induce different marginal incentives for

[22] The average loss due to implementing a suboptimal effort, $\widetilde{m} \times$ (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 test-optimal contract given an A/B test and the cost-minimizing contract that implements the same effort level, which of course, need not equal the benchmark-optimal effort level.

different agents. Second, different agents may respond differently to a change in their marginal incentives. Using data from an aggregate A/B test to infer agents' heterogeneous behavioral responses therefore requires imposing more structure on the problem. In this section, we show how to extend the conditions from Section III in a way that ensures that aggregate data from an A/B test is sufficient for solving the principal's problem, and we quantitatively explore the errors that arise when these extended conditions are not satisfied.

To this end, suppose that a share p_ϕ of agents has cost type $\phi \in \Phi$, where $p_\phi \geq 0$ and $\sum_\phi p_\phi = 1$.[23] Suppose further that the principal has access to what we refer to as an **aggregate A/B test**, $\overline{AB}(w^A, w^B) = (\bar{f}^A, \bar{f}^B)$, where $\bar{f}^A(x) = \sum_\phi p_\phi f(x \mid a_\phi(w^A))$, and, abusing notation slightly, $a_\phi(w)$ is the effort choice for a type-ϕ agent under contract w. The density $\bar{f}^B(x)$ is defined similarly. Define $\bar{a}(w) = \sum_\phi p_\phi a_\phi(w)$ to be the **mean effort under contract** w.

Throughout this section, we assume that the output distribution satisfies Condition 1, that is, $f(x \mid a) = g(x) + ah(x)$ for all a and for some $g(x)$ and $h(x)$ satisfying $\int g(x)\,dx = 1$ and $\int h(x)\,dx = 0$. We also assume that for each type $\phi \in \Phi$, the agent's effort-cost function c_ϕ satisfies Condition 2 for some $\varepsilon_\phi, \beta_\phi \geq 0$; that is, $c'_\phi(a) = e^{-\beta_\phi/\varepsilon_\phi} a^{1/\epsilon_\phi}$. Finally, we modify the principal's problem so that the optimal adjustment only has to make agents better off *on average* than the status quo contract.[24] That is, if we denote the principal's profit under contract w as

$$\pi(w) = \sum_{\phi \in \Phi} p_\phi m a_\phi(w) - \sum_{\phi \in \Phi} p_\phi \int w(x)\left[g(x) + a_\phi(w)h(x)\right] dx,$$

the principal's problem is to

$$\text{maximize}_w \pi(w) \ \text{ subject to } \sum_\phi p_\phi u_\phi(w) \geq \sum_\phi p_\phi u_\phi(w^A),$$

where

$$u_\phi(w) = \int v(w(x))\left[g(x) + a_\phi(w)h(x)\right] dx - c_\phi(a_\phi(w)).$$

We will first show how Condition 1 allows us to compute agents' marginal incentives using an aggregate A/B test, even if agents are heterogeneous. By Condition 1, we have $f(x \mid a_\phi(w)) = g(x) + a_\phi(w)h(x)$ for each ϕ and therefore, if we average over ϕ, we have $\bar{f}^k(x) = g(x) + \bar{a}(w^k)h(x)$ for $k \in \{A, B\}$. The function $h(x)$ therefore satisfies $h(x) = (\bar{f}^B(x) - \bar{f}^A(x))/(\bar{a}(w^B) - \bar{a}(w^A))$ for all x and can be computed using only information from an aggregate A/B test. Condition 1 also ensures that marginal incentives are independent of the agent's effort choice and therefore are common across agents for a given contract w; that is, $I(w) = \int v(w(x))h(x)\,dx$.

[23] The results in this section are prior free, so it is immaterial whether the principal knows p_ϕ.

[24] As in the main model, this constraint is motivated by the fact that contract changes are often viewed by employees with skepticism. So if it makes them better off on average, it is less likely that a critical mass will oppose it. Additionally, given aggregate data alone, the principal cannot evaluate how a contract change affects the expected payoff of each individual type.

Next, consider the procedure we outlined in Section III for how to use an A/B test to predict effort under contract w when agents are homogeneous, and denote this prediction by $\hat{a}(w)$; that is, using an aggregate A/B test, compute the arc elasticity

$$\bar{\bar{\varepsilon}} = \frac{\ln\bar{a}(w^A) - \ln\bar{a}(w^B)}{\ln I(w^A) - \ln I(w^B)},$$

and construct the prediction

$$\hat{a}(w) = \bar{a}(w^A)\left[I(w)/I(w^A)\right]^{\bar{\bar{\varepsilon}}}.$$

In other words, this procedure predicts that a contract that scales marginal incentives over the status quo contract by $I(w)/I(w^A)$ will scale mean output by $\left[I(w)/I(w^A)\right]^{\bar{\bar{\varepsilon}}}$.

The following result focuses on the case when all agents have the same elasticity; i.e., $\varepsilon_\phi = \varepsilon$ for all ϕ.[25]

PROPOSITION 4: *Suppose Conditions 1 and 2 are satisfied, and agents have the same elasticity of effort with respect to marginal incentives; that is,* $\varepsilon_\phi = \varepsilon$ *for all* ϕ*. Then this procedure produces the correct prediction (i.e.,* $\hat{a}(w) = \bar{a}(w)$ *for all* w*), and an aggregate A/B test suffices for solving the principal's problem.*

The first part of Proposition 4 shows that aggregate information can be used to construct correct predictions about how a heterogeneous workforce responds to a change in the contract. There are two key steps in the argument. First, as we described above, when Condition 1 holds, the agents' marginal incentives depend only on the contract they face and not directly on their effort. Given this property, different types all face exactly the same marginal incentives, and a given adjustment changes their marginal incentives in exactly the same way. Second, when Condition 2 holds and agents have the same elasticity of effort with respect to their marginal incentives, a given change in marginal incentives leads all agents to scale their effort by the same proportion. To establish the second result that an aggregate A/B test suffices to solve the principal's problem, the proof of Proposition 4 shows that calculating the principal's objective and the agents' mean utility depends only on having a correct prediction of the function $\bar{a}(\cdot)$.

We now discuss the case when agents differ in ε_ϕ. Given an aggregate A/B test, the principal's prediction for how mean output changes with w, $\hat{a}(w) = \bar{a}(w^A)\left[I(w)/I(w^A)\right]^{\bar{\bar{\varepsilon}}}$, will be incorrect. Different agent types will have different proportional responses to the change in marginal incentives, and so the actual mean output under contract w will be $\bar{a}(w) = \sum_\phi p_\phi a_\phi(w^A)\left[I(w)/I(w^A)\right]^{\varepsilon_\phi}$.

Figure 4 quantifies the resulting discrepancy. The left panel plots probability mass functions for three distributions over ε_ϕ. The distribution depicted by blue squares second-order-stochastically dominates the distribution depicted by green triangles, which in turn second-order-stochastically dominates the distribution depicted by red circles. The panel on the right plots the systematic prediction error that arises under each of these three distributions when the principal uses the A/B test comprising

[25] This form of heterogeneity has been assumed elsewhere, for example, by Brewer, Saez and Shephard (2010); DellaVigna and Pope (2018); and others.

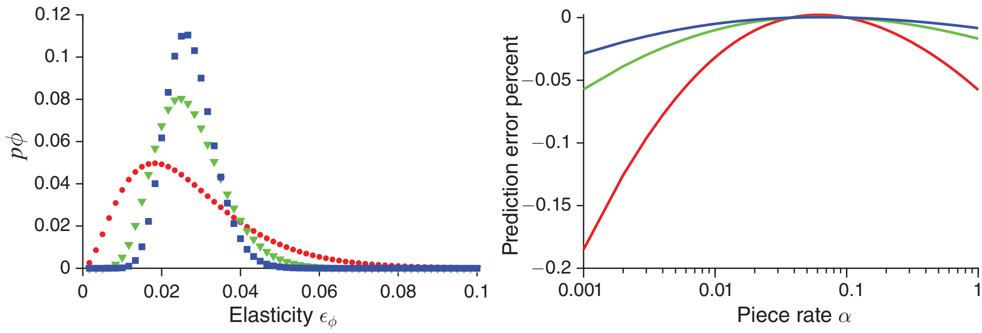

Figure 4

Notes: This figure illustrates for three different distributions over ϵ_ϕ, the prediction errors when the principal uses aggregate data from the A/B test comprising treatments $w^4(x)$ and $w^5(x)$ from Table 1 to make out-of-sample predictions for contracts of the form $w(x) = 100 + \alpha x$, where we vary α from 0.001 to 1. The left panel plots the probability mass functions for three distributions over ε_ϕ, and the right panel plots the prediction error that arises under each of these three distributions as a function of α.

treatments $w^4(x)$ and $w^5(x)$ from Table 1 and assumes coefficient of RRA $\rho = 0.3$. On the vertical axis, it plots $100\% \times (\hat{a}(w) - \bar{a}(w))/\bar{a}(w)$, and on the horizontal axis, it varies the slope α of a piece-rate contract between 0.001 and 1.[26]

The right panel of Figure 4 highlights several patterns. First, the prediction error is zero for contracts that induce the same marginal incentives as either the status quo contract or the test contract. Second, this error is positive (but small) for contracts that induce marginal incentives in between those induced by the status quo and test contracts, and it is negative for contracts with marginal incentives outside this range. Third, this error is larger in magnitude the more agents vary in ε_ϕ. Fourth, it is also larger in magnitude when we predict effort under contracts that are farther away from the status quo and test contracts (in the sense that they induce much higher or much lower marginal incentives). And finally, this error is relatively small in magnitude: it is less than 0.2 percent for the most disperse distribution.

We conclude this section by discussing the consequences of ignoring heterogeneity in the agents' preferences over money. Toward this goal, suppose Conditions 1 and 2 are satisfied, and agents have CRRA preferences over money, but different types have different coefficients of RRA. Because the marginal incentives generated by any given contract depend on the agent's utility function, the principal will miscalculate them if she ignores any underlying heterogeneity. Her effort predictions will therefore be biased.

[26] We construct these probability mass functions as follows: First, we assume $\epsilon_\phi \sim Gamma(\kappa, \theta)$, where $\kappa \in \{3, 10, 20\}$ corresponds to the probability mass function depicted by red circles, green triangles, and blue squares, respectively; and for each κ, the scale parameter θ is determined below. To compute the probability weights p_ϕ, we discretize the gamma distribution on the grid $\epsilon_\phi \in \{0, \Delta, 2\Delta, \dots\}$ for $\Delta = 10^{-3}$. Second, we compute $\bar{\epsilon}$ using the (aggregate) data from the A/B test, and we assume, first, that $a_\phi(w^B) = \bar{a}(w^B)$ for all ϕ, and second, that agents have CRRA preferences over money with coefficient 0.3. Next we compute as a function of θ, the effort of each type under the status quo contract $a_\phi(w^A) = a_\phi(w^B)\left[I(w^A)/I(w^B)\right]^{\epsilon_\phi}$. Then, we pick the parameter θ such that $\ln \bar{a}(w^A) = \bar{\epsilon} \ln\left(I(w^A)/I(w^B)\right)$, thus ensuring that $\bar{\epsilon}$ is consistent with the distribution over elasticities. Finally, we note that in light of Proposition 4, any heterogeneity in β_ϕ can be ignored without loss of generality, and that the other A/B tests yield no larger prediction errors.

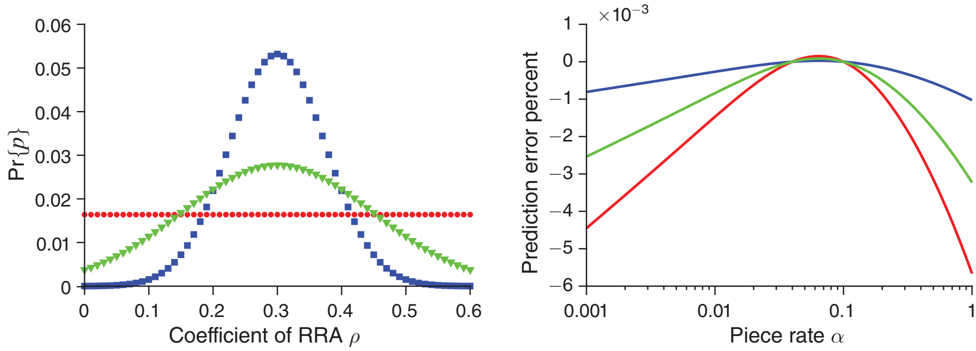

FIGURE 5

Notes: This figure illustrates for three different distributions over the agents' coefficient of RRA, the prediction errors when the principal uses the A/B test comprising treatments $w^4(x)$ and $w^5(x)$ from Table 1 and assumes agents have homogeneous elasticities and a common coefficient of RRA $\rho = 0.3$ to make out-of-sample predictions for contracts of the form $w(x) = 100 + \alpha x$, where we vary α from 0.001 to 1.

Figure 5 quantifies this bias. The left panel plots probability mass functions for three distributions over the coefficient of RRA. The panel on the right plots the systematic prediction error that arises under each of these three distributions when the principal uses the A/B test comprising treatments $w^4(x)$ and $w^5(x)$ from Table 1 and assumes agents have homogeneous elasticities and common coefficient of RRA $\rho = 0.3$. On the vertical axis, it plots the prediction error, and on the horizontal axis, it varies the slope α of a piece-rate contract between 0.001 and 1. Observe that in all cases, the prediction error is negligible.[27]

B. *Multidimensional Effort*

We now extend our main model to the case where the agent's action is multidimensional. For example, the agent might be selling different products; he might exert effort toward both quantity and the quality of his output; or he might be able to influence several aspects of his output distribution, for example, its mean and its variance. Two additional challenges arise when extending our methodology to accommodate multidimensional effort. First, effort along one dimension might affect the marginal costs of effort along other dimensions. Identifying these effort-substitution patterns requires additional test contracts in the local A/B test. Constructing the local A/B test from a set of test contracts presents a second challenge. In contrast to the one-dimensional case where effort could be normalized to be mean output, identifying the agent's effort vector from data on the output distribution requires additional a priori information about the nature of effort and may necessitate additional test contracts.

[27] We have assumed that the principal's estimate, $\rho = 0.3$, is unbiased; i.e., it is equal to the expectation over the agents' coefficients of RRA. If this is not the case, the prediction error will be larger. In this example, if the principal assumes a common coefficient of RRA $\rho = 0.1$ or $\rho = 0.5$ instead, the prediction error will remain below 1 percent.

To examine these issues, suppose the agent chooses a vector of actions $\mathbf{a} \in \mathbb{R}^M$, and a vector of performance measures $\mathbf{x} \in \mathbb{R}^N$ is realized according to the pdf $f(\cdot \,|\, \mathbf{a})$. The agent is paid according to a contract $w(\mathbf{x})$, and the cost of choosing effort vector $\mathbf{a}$ is $c(\mathbf{a})$, where c is increasing and convex. Given a contract w, the agent's utility is

$$u(w) = \max_{\mathbf{a}} \int v(w(\mathbf{x})) f(\mathbf{x} \,|\, \mathbf{a})\, d\mathbf{x} - c(\mathbf{a}),$$

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

$$\mathcal{D}u(w,t) = \int t v'(w) f\, d\mathbf{x},$$

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

$$\sum_{k=1}^{M} \left[c_{i,k} - \int v(w) f_{i,k}\, d\mathbf{x}\right] \mathcal{D}a_k(w,t) = \int t v(w) f_i \, d\mathbf{x},$$

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

Given contract w and an adjustment t, let us define the **Hessian matrix A** to be an $M \times M$ symmetric matrix with elements $A_{i,k} = c_{i,k} - \int v(w) f_{i,k} d\mathbf{x}$. Note that this matrix does not depend on the adjustment t. Let us also define the **marginal-incentives matrix** under adjustment t to be the $M \times 1$ matrix $\mathbf{B}(t)$ with elements $B_i(t) = \int t v(w) f_i \, d\mathbf{x}$. We can then write the multidimensional analog of (2) as $\mathcal{D}\mathbf{a}(w,t) = \mathbf{A}^{-1}\mathbf{B}(t)$, where $\mathcal{D}\mathbf{a}(w,t)$ denotes the $M \times 1$ matrix with kth element $\mathcal{D}a_k(w,t)$.

Next, we turn to the principal's profits. Again, using the same approach as in Section II, adjusting a contract w in the direction t changes her profit according to the differential

$$\mathcal{D}\pi(w,t) = \sum_{i=1}^{M} \left[m_i - \int w(\mathbf{x}) f_i(\mathbf{x}|\mathbf{a}(w))\, d\mathbf{x}\right] \mathcal{D}a_i(w,t) - \int t(\mathbf{x}) f(\mathbf{x}|\mathbf{a}(w))\, d\mathbf{x},$$

where notice that we are allowing the principal to place different values m_i on different dimensions of effort. Given a status quo contract w^A, the principal solves

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

Turning to the information required for solving the principal's problem, let us denote a **local A/B test with** K **test contracts** $w^{B_1}, \ldots, w^{B_K}$ by $LAB(w^A, w^{B_1}, \ldots, w^{B_K}) = (f^A, \nabla f^A, \mathcal{D}\mathbf{a}(w^A, w^{B_1}), \ldots, \mathcal{D}\mathbf{a}(w^A, w^{B_K}))$, where ∇f^A is an $M \times 1$ matrix with ith element $f_i(\mathbf{x}|\mathbf{a}(w^A))$. We will say that test contracts $w^{B_1}, \ldots, w^{B_K}$ are **informative and independent** if $\mathcal{D}\mathbf{a}(w^A, w^{B_k}) \neq 0$ for all k, and $\mathcal{D}\mathbf{a}(w^A, w^{B_1}), \ldots, \mathcal{D}\mathbf{a}(w^A, w^{B_K})$ are linearly independent.

Recall that Proposition 1 shows that, for the unidimensional effort case, a local A/B test reveals f^A, f^A_a, and enables the principal to compute how the agent's marginal incentives and utility change for any adjustment t. By the same logic, when $M \geq 2$, knowledge of f^A and ∇f^A suffices for constructing the agent's marginal incentives matrix $\mathbf{B}(t)$ and computing $\mathcal{D}u(w^A, t)$ for any t.

When $M = 1$, the agent's Hessian matrix $\mathbf{A}$ is a singleton, and Proposition 1 shows that it can be identified with a single test contract. When $M \geq 2$, the agent's Hessian matrix contains $M(M+1)/2$ distinct elements, as it is symmetric. These elements cannot all be inferred from a local A/B test with one test contract, but knowledge of $\mathcal{D}\mathbf{a}(w,t)$ for a *particular* adjustment t, together with f^A and ∇f^A generates M equations of the form $\mathcal{D}\mathbf{a}(w,t) = \mathbf{A}^{-1}\mathbf{B}(t)$. The matrix $\mathbf{A}$ can therefore be identified as long as the principal knows $\mathcal{D}\mathbf{a}(w,t)$ for at least $\lceil (M+1)/2 \rceil$ informative and independent test contracts. Given an estimate for $\mathbf{A}$, one can then compute $\mathcal{D}\mathbf{a}(w^A, t)$ and therefore $\mathcal{D}\pi(w^A, t)$ for every t. Therefore, a local A/B test with $K = \lceil (M+1)/2 \rceil$ informative and independent test contracts provides all the information needed to solve the principal's problem.

We now address the second challenge that arises when effort is multidimensional: *constructing* a local A/B test. In the unidimensional effort case, constructing a local A/B test from output data is straightforward. There, it is without loss of generality to normalize effort so that $E[x|a] = a$, so that by observing the output distribution for a given contract, the principal can infer the chosen effort. Then, given contracts w^A and w^B, the principal can construct $\mathcal{D}a(w^A, w^B) \approx a^B - a^A$, and $f^A_a(x) \approx [f^B(x) - f^A(x)]/(a^B - a^A)$.

When effort is multidimensional, using output data from K test contracts to construct a local A/B test requires a priori information on the nature of effort, and it may also put a lower bound on how many test contracts are required. To illustrate the first point, define the function $G : \mathbb{R}^M \rightarrow \mathbb{R}^N$ such that $G_i(\mathbf{a}) = E[x_i \,|\, \mathbf{a}]$ for each i. If G is invertible and known by the principal, then observing $E[\mathbf{x} \,|\, \mathbf{a}(w)]$ for some contract w suffices to infer $\mathbf{a}(w)$, and therefore the principal can use output data to construct $\mathcal{D}\mathbf{a}(w^A, w^{B_k})$ for each k.

The assumptions that G is invertible and known by the principal are restrictive but capture many potential settings of interest. For example, suppose the agent is a salesperson selling M different products, and his effort a_i affects only the distribution of his sales x_i of product i. Then we can let $M = N$, and $G_i(\mathbf{a}) = a_i$ is once again a normalization. As another example, suppose output y is one-dimensional, but a_1 influences mean output and a_2 the variance of output. This setting can be captured by setting $N = 2$ and letting $x_1 = y$, $x_2 = y^2$, $m_1 = m$, $m_2 = 0$, $G_1(\mathbf{a}) = a_1$, and $G_2(\mathbf{a}) = a_2 + a_1^2$. This example highlights that even when output is low-dimensional, the output distribution contains a lot of information that may be informative about the agent's choices.

Finally, to illustrate why constructing a local A/B test from output data may require additional test contracts, note that, given contracts w^A and w^{B_k}, we have for every $\mathbf{x}$,

$$f(\mathbf{x} \,|\, \mathbf{a}^{B_k}) - f(\mathbf{x} \,|\, \mathbf{a}^A) \approx \nabla f(\mathbf{x} \,|\, \mathbf{a}^A) \cdot \mathcal{D}\mathbf{a}(w^A, w^{B_k}).$$

Knowing $\mathcal{D}\mathbf{a}(w^A, w^{B_k})$ does not generally suffice to infer $\nabla f(\mathbf{x}|\mathbf{a}^A)$. To infer $\nabla f(\mathbf{x}|\mathbf{a}^A)$, one must solve a linear system with M unknowns, which means that *up to M* informative and independent test contracts may be required. Oftentimes, however, $\nabla f(\mathbf{x}|\mathbf{a}^A)$ can be identified with a single test contract: for example, if $M = N$ and f is separable so that each a_i determines only the distribution of x_i, then $\partial f^A / \partial a_i$ can be determined using the same identity as in the unidimensional case.

We conclude this section with a discussion of how these ideas can be applied non-locally. Recall from Section A that the treatment pairs (4,6), (5,6), and (5,7) generate similar mean output but starkly different output distributions. This is because subjects can adjust their efforts over time, suggesting their actions are multidimensional. In online Appendix A.C, we show how to extend the analysis to a setting in which subjects are allowed to choose the entire output distribution. As we shall see, however, it is important for the principal to take a stance on the nature and the dimensions of effort a priori, and moreover, the right kind of contract variation may be needed to learn about different dimensions of effort.

Finally, multiple test contracts are likely to prove useful in empirical settings such as those of Gibbs, Neckermann, and Siemroth (2017) and Hong et al. (2018), where agents exert effort toward both quantity and quality. To use multiple test contracts to derive optimal non-local adjustments in these settings, one would have to impose assumptions analogous to Conditions 1 and 2. For example, one might assume that output is separable and affine in each dimension; and the cost function has scale, elasticity, and cross-elasticity parameters. Each test contract provides two first-order conditions, and so to recover the unknown parameters, outcome data from three contracts would be needed.

VI. Discussion and Avenues for Future Research

What does a manager need to know to improve upon an existing contractual arrangement? We asked and answered this question in the context of the Holmström (1979) model of principal-agent relationships subject to pure moral hazard problems, we showed how A/B contracts can provide the relevant information, and we carried out an empirical proof of concept.

In the last forty years, contract theory has greatly extended its domain, but it has largely strayed away from the kinds of measurement issues that are important in practice. This paper just scratches the surface of what we hope can be a fruitful research agenda that combines theoretical insights with data to answer practical incentive-design questions. There are still important hurdles to overcome and many important directions to extend the analysis.

Our framework sidesteps both statistical error and approximation error. First, we assumed the principal has access to an infinitely large sample of output draws under each contract she has outcome data for. Understanding the limitations of smaller sample sizes is important for applications, especially in smaller firms.

Second, when we considered non-local adjustments, the conditions we imposed can be interpreted as an approximation to the true model, which may become worse when considering contracts farther away from the status quo contract. When this is the case, not all A/B contracts are equally informative, and questions of optimal A/B test design become more central (see, for example, Azevedo et. al. 2020).

Optimal A/B test design should be informed both by theories of approximation error and by empirical findings. In our empirical context, we found that homogeneous A/B contracts tend to lead to better performance than hybrid A/B contracts. And A/B contracts that themselves lead to large performance changes tend to lead to better performance than those that induce similar performance. We discussed the reasons for these differences at the end of Section A. Our analysis also sheds light on cases in which data from additional test contracts is needed—namely, if external incentives are important (or relatedly, one of the test contracts generates zero marginal incentives), or if effort is multidimensional.

Our framework also implicitly assumes that the outcome data given by an A/B test are generated by nonstrategic agents who are best responding to the contract they face. If agents know they are part of an experiment that will inform their future compensation, ratchet effects may reduce the informativeness of the A/B test. Similarly, if agents have other-regarding preferences (see, for example, Bandiera, Barankay, and Rasul 2011), then agents under one contract might react negatively to the knowledge that their coworkers face a different contract. In some settings, these distortions can be avoided altogether by appropriately choosing high-level features of the test contract. For example, ratchet effects can be ameliorated by using aggregate output data from many agents, since the resulting free-rider problem among agents during the test phase will tend to push each of them toward choosing an effort level that is a static response to the contract (Cardella and Depew 2018). Or, if agents' other-regarding preferences are determined at the team level, then assigning treatments at the team level, rather than the individual, can prevent negative reactions. In other settings, these kinds of considerations may be unavoidable and will therefore inform the design and informativeness of the experiment itself (see, for example, Liang and Madsen 2020 in the presence of strategic manipulation and Fehr, Powell, and Wilkening 2021 in the presence of negative reciprocity).

We showed how to extend our framework to accommodate several additional considerations that are not present in the canonical model, but there are many other important considerations that we did not incorporate. For example, in many environments, team production makes it hard to distinguish individual performance, and one agent's marginal incentives may depend on the effort choices of other agents. When this is the case, there may be value in putting different agents in the same team on different test contracts.

The last consideration that we will close with is that many workers are motivated through the use of long-term incentives arising from promotion systems or deferred compensation policies. In many models of dynamic incentives, an agent's marginal incentives are summarized by the sensitivity of their continuation payoffs to their current performance. A/B contracts can still be used to assess how agents respond to a change in today's marginal incentives, but to understand how best to adjust dynamic contracts, the principal would need additional information on how agents trade off today's compensation with their future career prospects in the firm.

REFERENCES

Azevedo, Eduardo M., Alex Deng, Jose Luis Montiel Olea, Justin Rao, and E. Glen Weyl. 2020. "A/B Testing with Fat Tails." *Journal of Political Economy* 128 (12): 4614–72.

Baily, Martin Neil. 1978. "Some Aspects of Optimal Unemployment Insurance." *Journal of Public Economics* 10 (3): 379–402.

Balbuzanov, Ivan, Jared Gars, and Emilia Tjernström. 2017. "Media and Motivation: The Effect of Performance Pay on Writers and Content." Unpublished.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2011. "Field Experiments with Firms." *Journal of Economic Perspectives* 25 (3): 63–82.

Barron, Daniel, George Georgiadis, and Jeroen Swinkels. 2020. "Optimal Contracts with a Risk-Taking Agent." *Theoretical Economics* 15 (2): 715–61.

Bénabou, Roland, and Jean Tirole. 2002. "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics* 117 (3): 871–915.

Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3): 489–520.

Bénabou, Roland, and Jean Tirole. 2006. "Belief in a Just World and Redistributive Politics." *Quarterly Journal of Economics* 121 (2): 699–746.

Brewer, Mike, Emmanuel Saez, and Andrew Shephard. 2010. "Means-Testing and Tax Rates on Earnings." In *Dimensions of Tax Design: The Mirrlees Review*, edited by Stuart Adam et al., 90–173. Oxford: Oxford University Press.

Cardella, Eric, and Briggs Depew. 2018. "Output Restriction and the Ratchet Effect: Evidence from a Real-Effort Work Task." *Games and Economic Behavior* 107: 182–202.

Carroll, Gabriel. 2015. "Robustness and Linear Contracts." *American Economic Review* 105 (2): 536–63.

Chade, Hector, and Jeroen Swinkels. 2019. "Disentangling Moral Hazard and Adverse Selection." Unpublished.

Chetty, Raj. 2006. "A General Formula for the Optimal Level of Social Insurance." *Journal of Public Economics* 90 (10–11): 1879–1901.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge between Structural and Reduced-Form Methods." *Annual Review of Economics* 1 (1): 451–87.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99 (4): 1145–77.

Chung, Doug J., Byungyeon Kim, and Niladri B. Syam. 2020. *A Practical Approach to Sales Compensation: What Do We Know Now? What Should We Know in the Future?* Norwell, MA: Now Publishers.

DellaVigna, Stefano. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature* 47 (2): 315–72.

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69.

DellaVigna, Stefano, and Devin Pope. 2021. "Stability of Experimental Results: Forecasts and Evidence." Unpublished.

d'Haultfoeuille, Xavier, and Philippe Fevrier. 2020. "The Provision of Wage Incentives: A Structural Estimation Using Contracts Variation." *Quantitative Economics* 11 (1): 349–97.

Einav, Liran, Amy Finkelstein, Iuliana Pascu, and Mark R. Cullen. 2012. "How General are Risk Preferences? Choices under Uncertainty in Different Domains." *American Economic Review* 102 (6): 2606–38.

Fehr, Ernst, and Lorenz Goette. 2007. "Do Workers Work More if Wages are High? Evidence from a Randomized Field Experiment." *American Economic Review* 97 (1): 298–317.

Fehr, Ernst, Michael Powell, and Tom Wilkening. 2021. "Behavioral Constraints on the Design of Subgame-Perfect Implementation Mechanisms." *American Economic Review* 111 (4): 1055–91.

Feldstein, Martin. 1999. "Tax Avoidance and the Deadweight Loss of the Income Tax." *Review of Economics and Statistics* 81 (4): 674–80.

Finkelstein, Amy, and Matthew J. Notowidigdo. 2019. "Take-up and Targeting: Experimental Evidence from SNAP." *Quarterly Journal of Economics* 134 (3): 1505–56.

Foarta, Dana, and Takuo Sugaya. 2021. "Wait-and-See or Step in? Dynamics of Interventions." *American Economic Journal: Microeconomics* 13 (1): 399–425.

Georgiadis, George, and Michael Powell. 2022. "Replication Data for: A/B Contracts." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E146062V1.

Gibbs, Michael, Susanne Neckermann, and Christoph Siemroth. 2017. "A Field Experiment in Motivating Employee Ideas." *Review of Economics and Statistics* 99 (4): 577–590.

Gottlieb, David, and Humberto Moreira. 2017. "Simple Contracts with Adverse Selection and Moral Hazard." Unpublished.

Grant, M., and S. Boyd. 2013. "CVX: Matlab Software for Disciplined Convex Programming." CVX Research. http://cvxr.com/cvx (accessed May 1, 2020).

Grossman, Sanford, and Oliver D. Hart. 1983. "An Analysis of the Principal-Agent Problem." *Econometrica* 51 (1): 7–45.

Guiteras, Raymond P., and B. Kelsey Jack. 2018. "Productivity in Piece-Rate Labor Markets: Evidence from Rural Malawi." *Journal of Development Economics* 131: 42–61.

Hansen, Bruce E. 2009. "Lecture Notes on Nonparametrics." Unpublished.

Harberger, A. C. 1964. "The Measurement of Waste." *American Economic Review* 54 (3): 58–76.

Holmström, Bengt. 1979. "Moral Hazard and Observability." *Bell Journal of Economics* 10 (1): 74–91.

Holmström, Bengt. 2017. "Pay for Performance and Beyond." *American Economic Review* 107 (7): 1753–77.

Holmström, Bengt. 1982. "Moral Hazard in Teams." *Bell Journal of Economics* 13 (2): 324–40.

Holmström, Bengt, and Paul Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives. *Econometrica* 55 (2): 303–28.

Holmström, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7: 24–52.

Hong, Fuhai, Tanjim Hossain, John A. List, and Migiwa Tanaka. 2018. "Testing the Theory of Multitasking: Evidence from a Natural Field Experiment in Chinese Factories." *International Economic Review* 59 (2): 511–36.

Innes, Robert D. 1990. "Limited Liability and Incentive Contracting with Ex-ante Action Choices." *Journal of Economic Theory* 52 (1): 45–67.

Ke, Rongzhu. 2008. "Identifying Contract Optimality Non-parametrically with Moral Hazard: First Order Approach and Statistical Inference." Unpublished.

Kleven, Henrik. 2020. "Sufficient Statistics Revisited." Unpublished.

Laffont, Jean-Jacques, and Jean Tirole. 1988. "The Dynamics of Incentive Contracts." *Econometrica* 56 (5): 1153–75.

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review* 90 (5): 1346–61.

Liang, Annie, and Erik Madsen. 2020. "Data Linkages and Incentives." Unpublished.

MacLeod, W. Bentley, and James M. Malcomson. 1988. "Reputation and Hierarchy in Dynamic Models of Employment." *Journal of Political Economy* 96 (4): 832–54.

Michaillat, Pascal, and Emmanuel Saez. 2019. "Optimal Public Expenditure with Inefficient Unemployment." *Review of Economic Studies* 86 (3): 1301–31.

Mirrlees, James A. 1976. "The Optimal Structure of Incentives and Authority within an Organization." *Bell Journal of Economics* 7 (1): 105–31.

Ortner, Juan, and Sylvain Chassang. 2018. "Making Corruption Harder: Asymmetric Information, Collusion, and Crime." *Journal of Political Economy* 126 (5): 2108–33.

Oyer, Paul. 2000. "A Theory of Sales Quotas with Limited Liability and Rent Sharing." *Journal of Labor Economics* 18 (3): 405–26.

Prendergast, Canice. 2015. "The Empirical Content of Pay-for-Performance." *Journal of Law, Economics, and Organization* 31 (2): 242–61.

Saez, Emmanuel. 2001. "Using Elasticities to Derive Optimal Income Tax Rates." *Review of Economic Studies* 68 (1): 205–29.

Shearer, Bruce. 2004. "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment." *Review of Economic Studies* 71 (2): 513–34.

Wilson, Robert B. 1993. *Nonlinear Pricing*. Oxford: Oxford University Press.