What Motivates Effort? Evidence and Expert Forecasts*

Stefano DellaVigna
UC Berkeley and NBER

Devin Pope
U Chicago and NBER

This version: June 6, 2016

Abstract

How much do different monetary and non-monetary motivators induce costly effort? Does the effectiveness line up with the expectations of researchers? We present the results of a large-scale real-effort experiment with 18 treatment arms. We compare the effect of three motivators: (i) standard incentives; (ii) behavioral factors like present bias, reference dependence, and social preferences; and (iii) non-monetary inducements from psychology. In addition, we elicit forecasts by behavioral experts regarding the effectiveness of the treatments, allowing us to compare results to expectations. We find that (i) monetary incentives work largely as expected, including a very low piece rate treatment which does not crowd out incentives; (ii) the evidence is partly consistent with standard behavioral models, including warm glow, though we do not find evidence of probability weighting; (iii) the psychological motivators are effective, but less so than incentives. We then compare the results to forecasts by 208 experts. On average, the experts anticipate several key features, like the effectiveness of psychological motivators. A sizeable share of experts, however, expects crowd-out, probability weighting, and pure altruism, counterfactually. This heterogeneity does not reflect field of training, as behavioral economists, standard economists, and psychologists make similar forecasts. Using a simple model, we back out key parameters for social preferences, time preferences, and reference dependence, comparing expert beliefs and experimental results.

*We thank Ned Augenblick, Dan Benjamin, Patrick Dejarnette, Jon de Quidt, David Laibson, John List, Benjamin Lockwood, Barbara Mellers, Don Moore, Sendhil Mullainathan, Jesse Shapiro, Uri Simonsohn, Erik Snowberg, Philipp Strack, Justin Sydnor, Dmitry Taubinsky, Richard Thaler, Mirco Tonin, Kevin Volpp, and the audiences at Bonn University, the London School of Economics, the Max Planck Institute in Bonn, UC Berkeley, the University of Philadelphia (Wharton), the 2016 JDM Preconference, and the Munich Conference on Behavioral Economics for useful comments. We also thank Thomas Graeber, Johannes Hermle, Jana Hofmeier, Lukas Kiessling, Tobias Raabe, Michael Sheldon, Patricia Sun, and Brian Wheaton for excellent research assistance. We are also very thankful to all the experts who took the time to contribute their forecasts. We are very grateful for support from the Alfred P. Sloan Foundation (award FP061020).

1 Introduction

Monetary incentives have long been used as a way to change behavior. More recently, policy-makers, researchers, and businesses have turned to behavioral economics and psychology for additional levers. An example of this trend is the formation of Behavioral Science Units within the UK and US governments with a mission to 'translate findings and methods from the social and behavioral sciences into improvements in Federal policies and programs.'

The behavioral and psychological literature is replete with findings and methods aimed at motivating people to work harder, limiting issues of self-control, and increasing pro-social acts, among other applications. However, a criticism of this literature is that, if anything, there are too many potential levers to change behavior, without a clear indication of their relative effectiveness. Different dependent variables and dissimilar participant samples can make direct comparisons of effect sizes across various studies difficult. In particular, given the disparate evidence, it is not clear whether even behavioral experts would be able to determine the relative effectiveness of various possible interventions in a particular setting.

In this paper, we design and run a large pre-registered experiment that allows us to compare the relative effectiveness of multiple treatments within one setting. We focus on a real-effort task with treatments including monetary incentives and non-monetary behavioral motivators. The treatments are, as much as possible, model-based, so as to relate the findings to the key underlying behavioral parameters in the models. These parameters should be easier to compare across settings, facilitating future tests of the effectiveness and reliability of behavioral models.

In addition to providing evidence on the efficacy of various treatments, we also elicit forecasts from academic experts on the effectiveness of the treatments. We argue that the collection of forecasts in advance of a study is a valuable step: it captures the beliefs of the research community on a topic, and also indicates in which direction, and how decisively, the results diverge from such beliefs. In our context, the forecasts are of interest because the experiment covers a range of models in behavioral economics, applied to a real-effort setting.

Turning to the details, we recruit 9,800 participants on Amazon Mechanical Turk (MTurk) – an online platform that allows researchers to post small tasks that require a human to perform. MTurk has become very popular for experimental research in marketing and psychology (Paolacci and Chandler, 2014) and is increasingly used in economics as well (e.g., Kuziemko, Norton, Saez, Stantcheva, 2015). The limited cost per subject and large available population on MTurk allow us to run 18 treatments with over 500 subjects in each treatment arm.

The task for the subjects is to alternately press the "a" and "b" buttons on their keyboards as quickly as possible for ten minutes. The 18 treatments attempt to motivate participant effort using i) standard incentives, ii) non-monetary psychological inducements, and iii) behavioral factors such as present bias, reference dependence, and social preferences.

We present three main findings about performance. First, monetary incentives have a strong

and monotonic motivating effect: compared to a treatment with no piece rate, performance is 33 percent higher with a 1-cent piece rate, and another 7 percent higher with a 10-cent piece rate. A simple model of costly effort estimated on these three benchmark treatments predicts performance very well not only in a fourth treatment with an intermediate (4-cent) piece rate, but also in a treatment with a very low (0.1-cent) piece rate that could be expected to crowd out motivation. Instead, effort in this very-low-pay treatment is 24 percent higher than with no piece rate, in line with the predictions of a model of effort for this size of incentive.

Second, non-monetary psychological inducements are moderately effective in motivating the workers. The three treatments increase effort compared to the no-pay benchmark by 15 to 21 percent, a sizeable improvement especially given that it is achieved at no additional monetary cost. At the same time, these treatments are less effective than any of the treatments with monetary incentives, including the one with very low pay. Among the three interventions, two modelled on the social comparison literature and one on task significance (Grant, 2008), a Cialdini-type comparison (Cialdini et al., 2007) is the most effective.

Third, the results using behavioral factors are generally consistent with behavioral models of social preferences, time preferences, and reference dependence, but with important nuances. Treatments with a charitable giving component motivate workers in a way consistent with warm glow but not pure altruism: the effect on effort is the same whether the charity earns a piece-rate return of 1 cent or 10 cents. We also find some, though quantitatively small, evidence of a reciprocal gift-exchange response to a monetary 'gift'.

Turning to time preferences, treatments with payments delayed by 2 or 4 weeks induce less effort than treatments with immediate pay, for a given piece rate. However, the decay in effort is exponential, not hyperbolic, in the delay (although the confidence intervals of the estimates do not rule out present bias). This finding is consistent with recent evidence of no present bias on monetary payments (Andreoni and Sprenger, 2012), as opposed to on real effort (Augenblick, Niederle, and Sprenger, 2015).

Finally, we provide evidence on two key components of reference-dependent models: loss aversion and overweighting of small probabilities. Using a claw-back design (Hossain and List, 2012), we find a larger response to an incentive framed as a loss than as a gain. Probabilistic incentives as in Loewenstein, Brennan, and Volpp (2007), though, induce less effort than a deterministic incentive with the same expected value. This result is not consistent with over-weighting of small probabilities (assuming the value function is linear or moderately concave).

In the second stage of this project, we measure the beliefs of academic experts about the effectiveness of the treatments. This allows us to capture where the research community stands with respect to (an application of) standard behavioral models. It also allows us to measure the extent to which the effectiveness of the various treatments lines up with these expectations.

Specifically, we surveyed researchers in behavioral economics, experimental economics, and psychology, as well as some non-behavioral economists. We provided the experts with the

results of the 3 benchmark treatments with piece-rate variation to help them calibrate how responsive participant effort was to different levels of motivation in this task. We then ask them to forecast the effort participants exerted in the other 15 treatment conditions. Out of 314 experts contacted, 208 experts provided a complete set of forecasts. Our initial, broad selection of experts and the 66 percent rate ensure a good coverage of behavioral experts.

The experts anticipate correctly several results, and in particular the effectiveness of the psychological inducements. Strikingly, the average forecast ranks in the exact order the six treatments without private performance incentives: two social comparison treatments, a task significance treatment, the gift exchange treatment, and two charitable giving treatments.

At the same time, the experts mispredict certain features. The largest deviation between the average expert forecast and the actual results is for the very-low-pay treatment, where experts on average anticipate a 12 percent crowd out, while the evidence indicates no crowd out. In addition, while the experts predict very well the average effort in the charitable giving treatments, they expect higher effort when the charity earns a higher return; the effort is instead essentially identical in the two charitable treatments. The experts also overestimate the effectiveness of the gift exchange treatment by 7 percent.

Regarding the delayed-payout treatments, the experts predict a pattern of effort consistent with present bias, while the evidence is most consistent with exponential discounting. However, the difference between the average forecast and the worker effort is not statistically significant.

Regarding reference dependence, the experts expect the loss framing to have about the same effect as a gain framing with twice the incentives, consistent with the Tversky and Kahneman (1991) calibration. The evidence from MTurker effort is largely in line with this forecast, if somewhat imprecise. Turning to the probability weighting results, the experts on average overestimate the effect of the treatments with probabilistic piece rates.

We then present three findings on the heterogeneity of forecasts. First, we document that, perhaps surprisingly, the forecasts do not materially differ depending on the main field of the expert: behavioral economists, psychologists, laboratory experimenters and non-behavioral economists appear to share, on average, similar priors. Second, we consider the heterogeneity of forecasts by treatment. For some treatments, such as the ones with gain-loss or the ones with delayed payments, there is broad agreement on the predictions. In other treatments, like crowd out of incentives or probability weighting, there is significant disagreement, possibly reflecting a relatively smaller literature. Third, the treatments with higher heterogeneity in forecasts also have higher heterogeneity of MTurker effort. Thus, the disagreement among experts in these treatments may stem from genuine diversity in the strength of those behavioral motivators in the population.

In the final part of the paper, we exploit the tight link between the experimental design and the model to estimate the parameters based on the observed effort levels. Specifically, we estimate key parameters for models of social preferences, time preferences, and reference

dependence. In addition, we use the expert forecasts to estimate expert beliefs about these same parameters, under the assumption that the experts share a similar model of costly effort.

We employ two estimation procedures. In the first one, we use a minimum distance estimator with moments given by the average effort in each treatment. Specifically, the average effort in the three benchmark treatments pins down a two-parameter cost of effort function (which we specified in a pre-analysis plan) and a motivation parameter. With these parameters set, performance in the other treatments is a simple function of the behavioral parameters. The advantage of this procedure is that the experts, in principle, could also estimate this model, since they observe the average effort for the benchmark treatments before making the forecasts.

A disadvantage of this procedure is that it assumes, counterfactually, no heterogeneity in effort within a treatment. In a second procedure, we allow for heterogeneity in the marginal cost of effort and estimate the model on individual effort data using non-linear least squares.[1]

The results for the two procedures are similar. With respect to social preferences, the effort in the charitable giving treatments supports a warm glow model, with no role for pure altruism. The median expert, instead, expects a pure altruism parameter $\alpha = .07$, with no warm glow. Regarding the time preferences, the median expert expects a β of 0.76, in line with estimates in the literature, while the point estimate for β from the MTurker effort (while noisy) is around 1. On reference dependence, assuming a value function calibrated as in Tversky and Kahneman (1992), we find *under*weighting of small probabilities, while the median expert expects (modest) overweighting. If we jointly estimate the curvature as well, the data can accommodate probability weighting, but for unrealistic values of curvature. Finally, we back out the loss aversion parameter using a linear approximation.

We explore complementary findings on expert forecasts in a companion paper (DellaVigna and Pope, 2016). We present different measures of expert accuracy, comparing individual forecasts with the average forecast. We also consider determinants of expert accuracy and compare the predictions of academic experts to those of other groups of forecasters: PhDs, undergraduates, MBAs, and MTurkers. Finally, we examine beliefs of experts about their own expertise and the expertise of others. Thus, the companion paper focuses on what makes a good forecaster, while this paper is focused on behavioral motivators and the extent to which experts on average anticipate the effects of the various treatments.

Our findings relate to a vast literature on behavioral motivators. While we cannot cite all related papers, our treatments relate to the literature on pro-social motivation (Andreoni, 1989 and 1990), crowd-out (Gneezy and Rustichini, 2000), present-bias (Laibson, 1997; O'Donoghue and Rabin, 1999), and reference dependence (Kahneman and Tversky, 1979; Koszegi and Rabin, 2006), among others. Several of our treatments have parallels in the literature, such as Imas (2014) and Tonin and Vlassopoulos (2015) on real effort and charitable giving. Indeed,

[1] We estimate the model on the data rounded into 100-point bins, allowing us to use the first-order condition of effort and thus the non-linear least squares estimation.

it has been our intent to largely build on existing studies and integrate them in a common setting. Two main features set our study apart. First, we consider all the above motivators and behavioral models in one common environment, allowing us to measure the relative effectiveness and test models, holding the setting constant. Second, we collect expert forecasts enabling us to compare the effectiveness of behavioral interventions with the expectations.

The emphasis on expert forecasts ties this paper to a small literature on forecasts of research results.[2] An early example within economics[3] is a competition among laboratory experimenters to forecast the result of a pre-designed laboratory experiment using learning models trained on data (Erev et al., 2010). We instead examine the ability of experts to make quick, intuitive forecasts, of the type done in an informal consulting, advising, or mentoring session.

A few recent economics papers include forecasts on a smaller scale. Coffman and Niehaus (2014) survey 7 experts on persuasion, while Sanders, Mitchell, and Chonaire (2015) ask 25 faculty and students from two universities questions on 15 select experiments run by the UK Nudge Unit. Groh, Krishnan, McKenzie and Vishwanath (2015) elicit forecasts on an RCT from audiences of 4 academic presentations. These complementary efforts suggest the need for a more systematic collection of expert beliefs about research findings.[4]

Our paper also relates to recent work on replication in psychology and experimental economics, including the use of prediction markets to capture beliefs about the replicability of experimental findings (Dreber et al., 2015 and Camerer et al., 2016). We emphasize the complementarity, as our study examines a novel real-effort experiment building on behavioral models, while the Science Prediction Market concerns the exact replication of existing protocols.

Our paper also adds to a growing literature on *structural behavioral economics* (Laibson, Repetto, and Tobacman, 2007; Conlin, O'Donoghue, and Vogelsang, 2007; DellaVigna, Malmendier, and List, 2012; Barseghyan, Molinari, O'Donoghue, and Teitelbaum, 2013; DellaVigna, Malmendier, List, and Rao, 2015). A unique feature is that we compare structural estimates of key behavioral parameters in the data to the parallel beliefs of experts.

The paper proceeds as follows. In Section 2 we motivate the treatments in light of a simple costly-effort model, and in Section 3 we present the design of the task and of the expert survey. We present the results of the treatments in Section 4 and the evidence on forecasts in Section 5. In Section 6 we derive the implied behavioral parameters and in Section 7 we conclude.

[2]There is a larger literature on forecasting about topics other than research results, e.g., the Good Judgment Project on national security (Tetlock, 2010 and Tetlock and Gardner, 2015). Several surveys, like the IGM Economic Expert panel, elicit opinions of experts about economic variables, such as inflation or stock returns.

[3]A famous case outside economics is the "GeneSweep" betting pool started in 2000 that collected bets about the number of human genes in DNA, which was concurrently being sequenced (Pennisi, 2003).

[4]Banerjee, Chassang, and Snowberg (2016) provide a framework on related issues of optimal experimentation.

2 Treatments and Model

In this section we motivate the 18 treatments in the experiment (Table 1) in light of a simple model of worker effort. As we will describe in more detail in the next section, the MTurk workers have ten minutes to complete a real-effort task (pressing a-b keys), with differences across the treatments in incentives and behavioral motivators. The model of costly effort, which we used to design the experiment and is registered in the pre-analysis plan, ties the 18 treatments to key behavioral models, like present bias and reference dependence.

Piece Rates. The first four treatments involve variation in the piece rate received by experiment participants to push buttons. (The piece rate is in addition to the advertised compensation of a $1 flat fee for completing the task). In the first treatment subjects are paid no piece rate ('*Your score will not affect your payment in any way*'). In the next three treatments there is a piece rate at 1 cent ('*As a bonus, you will be paid an extra 1 cent for every 100 points that you score*'), 10 cents ('*As a bonus, you will be paid an extra 10 cents for every 100 points that you score*'), and 4 cents ('*As a bonus, you will be paid an extra 4 cents for every 100 points that you score*'). The 1-cent piece rate per 100 points is equivalent to an average extra 15-25 cents, which is a sizeable pay increase for a 10-minute task in MTurk. The 4-cent piece rate and, especially, the 10-cent piece rate represent substantial payment increases by MTurk standards. These stated piece rates are the only differences across the treatments.

The 0-cent, 1-cent, and 10-cent treatments provide evidence on the responsiveness of effort to incentives for this particular task. As such, we provide the results for these benchmark treatments to the expert forecasters so as to facilitate their forecasts of the other treatments. Later, we use the results for these treatments to estimate a simple model of costly effort and thus back out the behavioral parameters.

Formally, we assume that participants in the experiment maximize the return from effort e net of the cost of effort. Let e denote the number of points (that is, alternating a-b presses). For each point e, the individual receives a piece-rate p as well as an intrinsic reward, $s > 0$. We interpret s as capturing in reduced form intrinsic motivation: workers derive utility from their effort. This specification captures, in reduced form, a norm or sense of duty to put in effort for an employer, or gratitude for the $1 flat payment for the 10-minute task. The presence of intrinsic motivation is important because otherwise, for $s = 0$, effort would equal zero in the no-piece rate treatment, counterfactually.

We assume a cost of effort function $c(e)$ which satisfies $c'(e) > 0$ and $c''(e) > 0$ for all $e > 0$. The cost of effort is assumed to be convex given the 10-minute time limit. Assuming risk-neutrality, an individual solves

$$\max_{e \geq 0} (s + p)e - c(e), \tag{1}$$

leading to the solution (when interior) $e^* = c'^{-1}(s + p)$. Optimal effort e^* is increasing in the

piece rate p and in the intrinsic motivation s. We consider two special cases for the cost function, discussed further in DellaVigna, List, Malmendier, and Rao (2015). The first function, which we pre-registered, is the power cost function $c(e) = ke^{1+\gamma}/(1+\gamma)$, characterized by a constant elasticity of effort $1/\gamma$ with respect to the value of effort.[5] Under this assumption, we obtain

$$e^* = \left(\frac{s+p}{k}\right)^{1/\gamma}. \tag{2}$$

A plausible alternative is that the elasticity decreases as effort increases. A function with this feature is the exponential cost function, $C(e) = k\exp(\gamma e)/\gamma$, leading to solution[6]

$$e^* = \frac{1}{\gamma}\log\left(\frac{s+p}{k}\right). \tag{3}$$

Under either function, the solution for effort has three unknowns, s, k, and γ which we can back out from the observed effort at different piece rates, as we do in Sections 4 and 6.

Figure 1 illustrates the model. For a given marginal cost curve $c'(e)$ (black line), changes in piece rate p shift the marginal benefit curve $s+p$, plotted for two levels of piece rate p. The optimal effort $e^*(p)$ is at the intersection of the marginal cost and marginal benefit curves.

We stress two key simplifying assumptions. First, we assume that the workers are homogeneous, implying (counterfactually) that they would all make the same effort choice in a given treatment. Second, even though the piece rate is earned after a discrete number of points (100 points, or 1,000 points below), we assume that it is earned continuously so as to apply the first-order conditions. We make these restrictive assumptions to ensure the model is simple enough to be estimated using just the three benchmark moments which the experts observe. In Section 6 we present an alternative estimation method which relaxes these assumptions.

Pay-Enough or Don't Pay at All. Motivated by the literature on motivational crowd-out (Deci, 1971), we design a treatment with very low pay as in Gneezy and Rustichini (2000): "*As a bonus, you will be paid an extra 1 cent for every 1,000 points that you score.*" Even by MTurk standards, earning an extra cent upon spending several minutes on effortful presses is a very limited reward for effort. Thus, it may be perceived as offensive and crowd out motivation. We model the treatment as corresponding to a piece rate $p = .001$, with a possible shift Δs_{CO} in motivation s:

$$e^*_{CO} = c'^{-1}(s + \Delta s_{CO} + p). \tag{4}$$

Social Preferences. The next two treatments involve charitable giving: "*As a bonus, the Red Cross charitable fund will be given 1 cent for every 100 points that you score*" and "*as*

[5] The first order condition is $k(e^*)^\gamma = v$ where v is the return per unit of effort (in our case equal to $s+p$). Thus, $e^* = (v/k)^{1/\gamma}$, and $\partial e^*/\partial v = (1/k\gamma) * (v/k)^{1/\gamma - 1}$. The elasticity is $\eta_{e,v} = (1/k\gamma) * (v/k)^{1/\gamma-1} v (v/k)^{-1/\gamma} = 1/\gamma$.

[6] The first order condition is $k\exp(\gamma e^*) = v$ where v is the return per unit of effort (in our case equal to $s+p$). Thus, $e^* = (1/\gamma)\log(v/k)$. Then $\partial e^*/\partial v = (1/\gamma) * (k/v)/k$ and the elasticity is $\eta_{e,v} = (1/\gamma v) * v/((1/\gamma)\log(v/k)) = 1/\log(v/k)$.

a bonus, the Red Cross charitable fund will be given 10 cents for every 100 points that you score." The rates correspond to the piece rates in the benchmark treatments, except that the recipient now is a charitable organization instead of the worker, similar to Imas (2014) and Tonin and Vlassopoulos (2015). The two treatments allow us to test a) how participants feel about money for a charity versus money for themselves and b) whether charitable giving in this setting conforms more closely to a pure altruism model à la Becker (1972), in which the individual takes into account the return to the charity, or to a warm glow model à la Andreoni (1989, 1990), in which the return to the charity may not matter. The optimal effort is

$$e^*_{CH} = c'^{-1}\left(s + \alpha p_{CH} + a * .01\right). \tag{5}$$

In the pure altruism model, the worker feels good for each dollar raised for the charity by exerting effort; as such, the altruism parameter α multiplies the return to the charity p_{CH} (equal to .01 or .10). In the warm glow model, the worker still feels good for helping the charity, but she does not pay attention to the actual return to the charity; she just receives a utility return a for each button press to capture a warm glow or social norm of generosity.[7] Without loss of generality, we multiply the warm glow parameter a by .01 (the return in the 1-cent treatment) to facilitate the comparison between the two social preference parameters.[8] Provided an estimate for s, k, and γ, the two charity treatments pin down α and a.

The final social preference treatment is a gift exchange treatment modelled upon Gneezy and List (2009): "*In appreciation to you for performing this task, you will be paid a bonus of 40 cents. Your score will not affect your payment in any way.*" In this treatment there is no piece rate, but the 'gift' may increase the interior motivation s by a factor Δs_{GE} reflecting reciprocity towards the employer[9]. Thus, the gift exchange effort equals

$$e^*_{GE} = c'^{-1}\left(s + \Delta s_{GE}\right). \tag{6}$$

Gneezy and List (2009) finds a significant initial increase in effort in response to a monetary 'gift', while some follow-up papers (including Kube, Marechal, and Puppe, 2013, Esteves-Sorenson, 2015 and DellaVigna, List, Malmendier, and Rao, 2015) do not find a significant impact of a monetary 'gift' on effort.

Time Preferences. Next, we have two discounting treatments: "*As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account two weeks from today.*" and "*As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account four weeks from today.*" The piece rate is 1 cent as in a benchmark treatment, but the payment is delayed from nearly

[7] This warm glow specification, which is parallel to DellaVigna et al. (2015) is not part of the pre-registration.

[8] Without rescaling, the estimates for a would be rescaled by 1/100.

[9] The experiments on gift exchange in the field are motivated by laboratory experiments on gift exchange and reciprocity (Fehr, Kirchsteiger, and Riedl, 1993; Fehr and Gachter, 2000).

immediate ('*within 24 hours*') in the benchmark treatments, to two or four weeks later. This corresponds to the commonly-used experimental questions to capture present bias (Laibson, 1997; O'Donoghue and Rabin, 1999; Frederick, Loewenstein, and O'Donoghue, 2002).

We model the treatments with delayed payment with a present-bias model:

$$e_t^* = c'^{-1}\left(s + \beta\delta^t p\right), \tag{7}$$

where β is the short-run impatience factor and δ is the long-run discounting factor. By comparing e_t^* in the discounting treatments to e^* in the piece rate treatments it is possible to back out the present bias parameter β and the (weekly) discounting factor δ.

An important caveat is that present-bias should apply to the utility of consumption and real effort, not to the monetary payments per se, since such payments can be consumed in different periods (Augenblick, Niederle, and Sprenger, 2015). Having said this, the elicitation of present bias using monetary payments is very common, with mixed evidence on the extent of present bias (e.g., Andreoni and Sprenger, 2012).

Reference Dependence. Next, we introduce treatments motivated by prospect theory (Kahneman and Tversky, 1979). A cornerstone of prospect theory is loss aversion, the idea that losses loom larger than gains (e.g., Barberis, 2013). To measure loss aversion, we lever gains and losses as incentives using a framing manipulation, as in Hossain and List (2012) and Fryer, Levitt, List, and Sadoff (2012). The first treatment promises a 40-cent bonus for achieving a threshold performance: "*As a bonus, you will be paid an extra 40 cents if you score at least 2,000 points. This bonus will be paid to your account within 24 hours.*" The second treatment promises a 40 cent bonus, but then stresses that this payment will be lost if the person does not attain a threshold score: "*As a bonus, you will be paid an extra 40 cents. This bonus will be paid to your account within 24 hours. However, you will lose this bonus (it will not be placed in your account) unless you score at least 2,000 points.*" The payoffs are formally equivalent in the two cases, but the framing of the bonus differs. A third treatment is also on the gain side, for a larger 80-cent payment: "*As a bonus, you will be paid an extra 80 cents if you score at least 2,000 points. This bonus will be paid to your account within 24 hours.*"

For the gain treatments, subjects can earn payment G ($0.40 or $0.80) if they exceed a target performance T. Following the Koszegi-Rabin (2006) gain-loss notation (but with a reference point given by the status quo), the decision-maker maximizes

$$\begin{aligned} &\max_{e\geq 0} se + \mathbf{1}_{\{e\geq T\}}G + \eta\left(\mathbf{1}_{\{e\geq T\}}G - 0\right) - c\left(e\right) \text{ or} \\ &\max_{e\geq 0} se + \mathbf{1}_{\{e\geq T\}}G\left(1+\eta\right) - c\left(e\right) \end{aligned} \tag{8}$$

The first term, $se + \mathbf{1}_{\{e\geq T\}}G$, captures the 'consumption' utility, while the second term, $\eta(\mathbf{1}_{\{e\geq T\}}G - 0)$, captures the gain utility relative to the reference point of no bonus; the weight on gain utility, η, is often parametrized at 1. In the loss treatment, the decision-maker

takes bonus G as reference point and thus maximizes

$$\max_{e\geq 0} se + \mathbf{1}_{\{e\geq T\}} G + \eta\lambda\left(0 - \mathbf{1}_{\{e<T\}} G\right) - c\left(e\right) \text{ or}$$
$$\max_{e\geq 0} se + \mathbf{1}_{\{e\geq T\}} G\left(1+\eta\lambda\right) - \eta\lambda G - c\left(e\right). \tag{9}$$

Conditions (8) and (9) lead to the same solution for $\lambda = 1$, but with $\lambda > 1$ (loss aversion) effort is higher in the loss treatment. Indeed, Hossain and List (2012) and Fryer, Levitt, List, and Sadoff (2012) find that the loss treatment is more effective in motivating effort.

The gain condition for $G = \$0.80$ has the purpose of benchmarking loss aversion.[10] As conditions (8) and (9) show, the incentive to reach the threshold T is $\left(1+\eta\right)G$ in the gain condition versus $\left(1+\lambda\eta\right)G$ in the loss condition. Thus, without deriving the optimal effort, we can compare effort $e_{G.80}$ in the 80-cent gain condition and effort $e_{L.40}$ in the 40-cent loss condition. The two efforts are the same if $.8\left(1+\eta\right) = .4\left(1+\eta\lambda\right)$ or $\lambda = \left(1+2\eta\right)/\eta$, or $\lambda = 3$ for the common assumption $\eta = 1$. In the Koszegi-Rabin (2006) notation, a loss aversion λ of 3 doubles the response to incentives on the loss side versus on the gain side, and is thus equivalent to the parametrization of $\lambda = 2$ in the Kahneman and Tversky (1979) notation.

In addition to loss aversion, a key component of prospect theory is probability weighting: probabilities are transformed with a probability weighting function $\pi\left(P\right)$ which overweights small probabilities and underweights large probabilities (e.g., Prelec, 1998 and Wu and Gonzalez, 1996). This motivates two treatments with stochastic piece rates, with expected incentives equal to the 1-cent benchmark treatment: "*As a bonus, you will have a 1% chance of being paid an extra $1 for every 100 points that you score. One out of every 100 participants who perform this task will be randomly chosen to be paid this reward.*" and "*As a bonus, you will have a 50% chance of being paid an extra 2 cents for every 100 points that you score. One out of two participants who perform this task will be randomly chosen to be paid this reward.*"

In these treatments, the subjects earn piece rate p with probability P, and no piece rate otherwise. The parameters p and P are chosen such that $p * P = 0.01$, the piece rate in the 1-cent benchmark treatment. The utility maximization is

$$\max_{e\geq 0} se + \pi\left(P\right)u\left(p\right)e - c\left(e\right),$$

where $u\left(p\right)$ is the (possibly concave) utility of payment with $u\left(0\right) = 0$, and $\pi\left(P\right)$ is the probability weighting. The number of button-presses is given by

$$e^*_{PW,P} = c'^{-1}\left(s + \pi(P)u\left(p\right)\right). \tag{10}$$

A probability weighting function with prospect theory features implies $\pi(0.01) \gg 0.01$ and $\pi(0.5) \approx 0.5$. Thus, for $u\left(p\right)$ approximately linear, effort will be highest in the condition with

[10] To our knowledge, this is the first paper to propose this third condition, which allows for a simple measure of the loss aversion parameter λ.

.01 probability of a \$1 piece rate: $e^*_{PW,P=.01} > e^*_{PW,P=.5} \approx e^*_{.01}$. Conversely, with no probability weighting and concave utility, the order is reversed: $e^*_{PW,P=.01} < e^*_{PW,P=.5} < e^*_{.01}$.

Psychology-based Treatments. A classical literature in psychology recognizes that human motivation is based to some degree on social comparisons (e.g., Maslow, 1943). In particular, Robert Cialdini has used comparisons to the achievements of others to induce motivation (e.g., Goldstein, Cialdini, and Griskevicius, 2008). In the ideal implementation, we would have informed the workers that a large majority of participants attain a high threshold (such as 2,000 points). Given that we wanted to only report truthful messages, we opted for: "*Your score will not affect your payment in any way. Previously, many participants were able to score more than 2,000 points.*"

A second social-comparison treatment levers the competitiveness of humans (e.g. Frank, 1984 within economics): "*Your score will not affect your payment in any way. After you play, we will show you how well you did relative to other participants.*" Recent papers in economics find that comparisons with others, even in the absence of monetary benefits, impacts productivity (Bandiera, Barankay, Rasul, 2013; Ashraf, Bandiera, Jack, 2012).

The final manipulation is based on the influential literature in psychology on task significance (Grant, 2008): workers work harder when they are informed about the significance of their job. Within our setting, we inform people that "*Your score will not affect your payment in any way. We are interested in how fast people choose to press digits and we would like you to do your very best. So please try as hard as you can.*"

We model these psychological treatments as in (6) with a shift Δs in the intrinsic reward.

3 Experiment and Survey Design

Design Logic. We designed the experiment with a dual purpose in mind. The first purpose is to obtain broad evidence on motivators of effort, comparing behavioral motivators to incentives. From this perspective, we wanted our treatments to cover a large range of behavioral factors, including present-biased preferences, reference dependence, and social preferences, three cornerstones of behavioral economics (e.g., Rabin, 1998; DellaVigna, 2009; Koszegi, 2014). We also wanted to include other motivators borrowed more directly from psychology.

The second purpose is to examine how experts forecast the impact of the various motivators. From this stand-point, we had five desiderata: (i) the experiment should have multiple treatments, to make the forecasting more informative; (ii) the sample size for each treatment had to be large enough to limit the role for sampling variation, since we did not want the experts to worry about the precision of the estimates; (iii) the differences in treatments had to be explained concisely and effectively, to give experts the best chance to grasp the design; (iv) the results should be available soon enough, so that the experts could receive timely feedback; and (v) the treatments and forecasting procedure should be disclosed to avoid the perception

that the experiments were selected on some criterion, i.e., ones with counterintuitive results.

After considering several options, we settled on a between-subject real-effort experiment run on Amazon Mechanical Turk (MTurk), varying the behavioral motivators across arms. MTurk is an online platform that allows researchers and businesses to post small tasks (referred to as HITs) that require a human to perform. Potential workers can browse the set of postings and choose to complete any task for the amount of money offered. MTurk has become very popular for experimental research in marketing and psychology (Paolacci and Chandler, 2014) and is also used increasingly in economics, for example for the study of preferences about redistribution (Kuziemko, Norton, Saez, Stantcheva, 2015).

The limited cost per subject and large available population on MTurk allow us to run several treatments, each with a large sample size, achieving goals (i) and (ii). Furthermore, the MTurk setting allows for a simple and transparent design (goal (iii)): the experts can sample the task and can easily compare the different treatments, since the instructions for the various treatments differ essentially in only one paragraph. The MTurk platform also ensures a speedy data collection effort (goal (iv)). Finally, we pre-registered both the experimental design and the survey, including a pre-analysis plan, to achieve goal (v).

3.1 Real-Effort Experiment

With the above framework in mind, we designed a simple real effort task on MTurk. The task involved alternating presses of 'a' and 'b' for 10 minutes, achieving a point for each a-b alternation, a task similar to those used in the literature (Amir and Ariely, 2008; Berger and Pope, 2011). While the task is not meaningful per se, it does have features that parallel clerical jobs: it involves repetition and it gets tiring, thus testing the motivation of the workers. It is also simple to explain to both subjects and experts.

Before the subjects do the task, they go through four screens. First is the recruiting screen on MTurk, specifying a \$1 pay for participating in an '*academic study regarding performance in a simple task.*'[11] This pay is quite generous given that average pay on MTurk is \$1.40 per hour according to Horton and Chilton (2010). Subjects that click through the link see a consent form which they have to click on. That takes them to a third page where they enter their MTurk ID and answer three demographic questions.

Following this page, the fourth screen provides instructions: '*On the next page you will play a simple button-pressing task. The object of this task is to alternately press the 'a' and 'b' buttons on your keyboard as quickly as possible for 10 minutes. Every time you successfully press the 'a' and then the 'b' button, you will receive a point. Note that points will only be rewarded when you alternate button pushes: just pressing the 'a' or 'b' button without alternating between the two will not result in points. Buttons must be pressed by hand only*

[11] We require that workers have an 80 percent approval rate and at least 50 approved previous tasks.

(key-bindings or automated button-pushing programs/scripts cannot be used) or the task will not be approved. Feel free to score as many points as you can.' Then, the participant sees a different final paragraph (bold and underlined) depending on the condition to which they were randomly assigned. For example, in the 10-cent treatment, the sentence reads '*As a bonus, you will be paid an extra 10 cents for every 100 points that you score. This bonus will be paid to your account within 24 hours.*' Table 1 reports the key content of this paragraph for all 18 treatments.[12] At the bottom of the page, subjects can try the task before proceeding.

On the fifth screen, subjects do the real task. As subjects press digits, the page shows a clock with a 10-minute countdown, the current points, and any earnings accumulated (depending on the condition) (Appendix Figures 1a-d). A sentence summarizes the condition for earning a bonus (if any) in that particular treatment. Thus, the 18 treatments differ in only three ways: the main paragraph on the fourth screen explaining the condition, the one-line reminder in the task screen, and the rate at which earnings (if any) accumulate on the task screen.

After the 10 minutes are over, the subjects are presented with the total points, the bonus payout (if any) and the total payout, and can leave a comment if they wish. The subjects are then thanked for their participation and given a validation code to redeem their earnings.

Pre-registration. We pre-registered the design of the experiment on the AEA RCT Registry as AEARCTR-0000714 ("*Response of Output to Varying Incentive Structures on Amazon Turk*"). Among the pre-registered details of the experiment, we specified the rule for the sample size. We aimed to recruit ideally 10,000 participants, and at least 5,000 participants based on a power study which is part of the pre-registration.[13] We ran the experiment for 3 weeks, at which point we had reached approximately 10,000 subjects.[14]

We also pre-specified the roles for inclusion in the sample. Quoting from the registration document, "*the final sample will exclude subjects that (i) do not complete the MTurk task within*

[12]For space reasons, in Table 1 we omit the sentence '*The bonus will be paid to your account within 24 hours.*' The sentence does not appear in the time discounting treatments.

[13]Quoting from the registration, "*based on 393 pilot participants, the standard deviation of points scored was around 740 and was similar across different treatments. Assuming that this is approximately the standard deviation of each treatment in the experiment and assuming a sample size of 5500 (305 per treatment), there is thus an 80% power to reject the null hypothesis of zero difference in average points between two treatments when the actual difference between the two treatments is 168.1 points. Assuming instead a sample size of 10,000 (555 per treatment), there is then an 80% power to reject the null hypothesis of zero difference when the actual difference is 124.6 points. Based on our pilot, different treatments can create differences in average points scored by as much as 400-500 points.*"

[14]The registration documents states '*The task will be kept open on Amazon Mechanical Turk until either (i) two weeks have passed or (ii) 10,000 subjects have completed the study, whichever comes first. If two weeks pass without 5500 subjects completing the task, then the task will be kept open (up to six weeks) until 5500 subjects are obtained.*' We deviated slightly from this rule by running the experiment for three weeks because we incorrectly thought that we registered a three-week duration. The deviation has minor impact as (i) 80 percent of subjects had been recruited by the end of week 2, and (ii) the authors did not monitor the experimental results during the three weeks (other than for the three benchmark conditions), thus removing the potential for selective stopping.

30 minutes of starting or (ii) exit then re-enter the task as a new subject (as these individuals might see multiple treatments) or (iii) score 4000 or more points (as we have learned from a pilot study of ˜300 participants that it is physically impossible to score more than 3500 points, so it is likely that these individuals are using bots)."

We debated whether to run the experiment before, or after, the collection of forecasts. We decided to run the experiment first so as to provide the forecasters with the results of three benchmark incentive treatments, thus conveying the curvature of the cost of effort function. At the same time, we wanted to ensure that there would be no leak of any results, before all the expert forecasts were collected. As such, as authors we did not have access to experimental results until after the survey collection. We designed a do file to monitor the sample size as well as results in the three benchmark treatments. A research assistant then ran this do file on the data and sent us daily updates which we monitored for potential data issues (such as the glitch with Qualtrics mentioned below). We accessed the results of the other treatments only at the end of September 2015, after the expert forecasts were collected.

Data Collection. The experiment ran for three weeks in May 2015. The initial sample consists of 12,838 MTurk workers who started our experimental task. Of these, 721 were dropped because of a technical problem with the survey over a several-hour period when the software program Qualtrics moved to a new server. Individuals during this time period experienced a malfunctioning of the counter that kept track of their scores. This sample exclusion, which we could not have anticipated, does not appear in the registration.

We then applied the three specified sample restrictions. We dropped (i) 48 workers for scoring above 4,000 points, (ii) 1,543 workers for failing to complete the experiment (for example, many participants only filled out the demographics portion of the experiment and were never assigned a treatment), and (iii) 364 workers for stopping the task and logging in again. (We stated in the instructions to the workers that they could not stop the task and log in again.) Two additional restrictions were added: we dropped 187 workers because their HIT was not approved for some reason (e.g. they did not have a valid MTurk ID) as well as 114 workers who never did a single button press. These participants may have experienced a technical malfunction or it may be that their results were not recorded for some reason.[15]

Summary Statistics. The final sample includes 9,861 subjects, about 550 per treatment. As Table 2 shows, the demographics of the recruited MTurk sample matches those of the US population for gender, and somewhat over-represents high-education groups and younger individuals. This is consistent with previous literature documenting that MTurkers are actually quite representative of the population of U.S. internet users (Ipeirotis, 2009; Ross et al., 2010; Paolacci et al., 2010) on characteristics such as age, socioeconomic status, and education levels.

[15]The two additional restrictions, which are immaterial for the results, were added before we analyzed the full data and were included in the pre-registration for the survey protocol AEARCTR-0000731 (see below).

3.2 Expert Survey

Survey. The survey of experts, registered as AEARCTR-0000731, is formatted with the platform Qualtrics and consists of two pages.[16] In the first and main page, the experts read a description of the task, including the exact wording seen by the MTurkers. The experts can experience the task by clicking on a link and see the screenshots viewed by the MTurk workers with another click. The experts are then informed of a prize that depends on the accuracy of their forecasts. "*Five people who complete this survey will be chosen at random to be paid [...] These five individuals will each receive $1,000 - (Mean Squared Error/200), where the mean squared error is the average of the squared differences between his/her answers and the actual scores.*" This incentive structure is incentive compatible: participants who minimize the sum of squared errors will indicate as their forecast the mean expected effort by treatment.[17]

The survey then displays the mean effort in the three benchmark treatments: no-piece rate, 1-cent, and 10-cent piece rate. The results are displayed using the same slider scale used for the other 15 treatments, except with a fixed scale. The experts then see a list of the remaining 15 treatments and create a forecast by moving the slider, or typing the forecast in a text box (though the latter method was not emphasized) (Appendix Figure 2). The experts can scroll back up on the page to review the instructions or the results of the benchmark treatments. In order to test for fatigue, we randomize across experts the order of the treatments (the only randomization in the survey). Namely, we designate six possible orders, always keeping related interventions together, in order to minimize the burden on the experts.

We decided ex ante the rule for the slider scale. We wanted the slider to include the values for all 18 treatments while at the same time minimizing the scope for confusion. Thus, we chose the minimum and maximum unit to be the closest multiple of 500 that is at least 200 units away from all treatment scores. A research assistant checked this rule against the results, leading to a score between 1,000 and 2,500.

In the second page of the survey we elicit a measure of confidence: the best guess of the number of forecasts expected to be within 100 points of the actual average effort in a treatment. We also elicit forecasts for other groups of experts, and we inquire whether the experts have used MTurk subjects in their research and whether they are aware of MTurk.

Experts. To form the group of behavioral experts, we form an initial list including: (i) authors of papers presented at the Stanford Institute of Theoretical Economics (SITE) in Psychology and Economics or in Experimental Economics from its inception until 2014 (for all years in which the program is online); (ii) participants of the Behavioral Economics Annual Meeting (BEAM) conferences from 2009 to 2014; (iii) individuals in the program committee and keynote speakers for the Behavioral Decision Research in Management Conference (BDRM)

[16] We provide further details on the survey in DellaVigna and Pope (2016).

[17] We avoided a tournament payout structure (paying the top 5 performers) which could have introduced risk-taking incentives; we pay instead five randomly drawn participants.

in 2010, 2012, and 2014; (iv) invitees to the Russell Sage Foundation 2014 Workshop on "Behavioral Labor Economics" and (v) a list of behavioral economists compiled by ideas42. We also add by hand a small number of additional experts. We then pare down this list of over 600 people to 314 researchers to whom at least one of the two authors had some connection.

On July 10 and 11, 2015 one of the two authors sent a personalized email to each of the 314 experts. The email provided a brief introduction to the project and task and informed the expert that an email with a unique link to the survey would be forthcoming from Qualtrics. An automated reminder email was sent about two weeks later to experts who had not yet completed the survey (and had not expressed a desire to opt out from communication). Finally, one of the authors followed up with a personalized email to the non-completers.[18]

Out of the 314 experts sent the survey, 213 completed it, for a participation rate of 68 percent. The main sample of 208 experts does not include 5 responses with missing forecasts for at least one of the 15 treatments. Table 3 shows the selection into response. Notice that the identity of the respondents is kept anonymous. On November 30, 2015, each expert received a personalized email with a link to a figure analogous to Figure 6 that also included their own forecasts. We also drew winners and distributed the prizes as promised.[19]

4 Effort By Treatment

4.1 Average Effort

Piece Rate Treatments. We start the analysis from the benchmark treatments which the experts had access to: the no-piece rate, 1-cent, and 10-cent treatments. As we discussed above, incentives have a powerful effect on effort in this task, raising performance from an average of 1,521 points (no piece rate) to 2,029 (1-cent piece rate) and 2,175 (10-cent). The standard error for the mean effort per treatment is around 30 points or less (Table 4), implying that differences across treatments larger than 85 points are statistically significant.

The effort for a fourth piece rate treatment, 4 cents for every 100 points, was assigned to the experts to forecast. This 4-cent treatment was designed not because it involved any behavioral factors, but to the contrary because it is possible to forecast effort in this treatment with a simple economic model. The effort in this treatment should be expected to lie between the 1-cent and the 10-cent treatments. Further, given a convex cost of effort the average output is likely to be closer to the 10-cent treatment than to the 1-cent treatment.

We use the 4-cent treatment as a partial validation for the model of effort presented in Section 2. Namely, we estimate the model using the effort in the 0-cent, 1-cent, and 10-cent

[18] We also collected forecasts from PhD students in economics, undergraduate students, MBA students, and a group of MTurk subjects. We analyze these results in DellaVigna and Pope (2016).

[19] Since the survey included PhDs, undergraduates, MBAs and MTurkers, two of the prizes went to the experts.

piece rate, and then predict out of sample the effort in the 4-cent treatment.

Model Estimation. For the estimation, we use a standard minimum distance estimator taking as moments the average effort in the 0-cent, 1-cent, and 10-cent treatments. The benchmark model which we pre-registered assumes a power cost function, leading to expression (2) for effort e^*. The expression has three unknown parameters which we estimate: the intrinsic motivation s, the cost curvature (and inverse of the elasticity) γ and the scaling parameter k. Hence, we are exactly identified with 3 moments and 3 parameters.

As Column 1 of Table 5 shows,[20] the cost of effort has a high estimated curvature ($\hat{\gamma} = 33$) and thus a low elasticity of 0.03. This is not surprising given that an order-of-magnitude increase in the piece rate (from 1 to 10 cents) increases effort by less than 10 percent. The estimated motivation $\hat{s}$ is very small: given the high curvature of the cost of effort function, even a small degree of motivation can reproduce the observed effort of 1,522 for zero piece rate.

How does the estimated model then predict out of sample the output for a 4-cent piece rate? Figure 2a displays the estimated marginal cost curve $c'(e) = \hat{k}e^{\hat{\gamma}}$ and the marginal benefit curves $\hat{s} + p$ for the different piece rates. The figure shows that, by design, the model perfectly fits the 0-cent, 1-cent, and 10-cent cases. The model then predicts a productivity for the 4-cent case of 2,116, very close to the actual effort of 2,132.

Arguably, the assumption of a constant elasticity embedded in the power cost function may not be appropriate. A function with declining elasticity, as discussed in Section 2, is the exponential cost function, $c(e) = k \exp(\gamma e)/\gamma$. Column 3 of Table 5 shows that, as with the power function, the intrinsic motivation s is estimated to be very small. The exponential function also perfectly fits the benchmark moments, and makes a similar prediction for the 4-cent treatment as the power function (Appendix Figure 3a). Further, allowing for heterogeneity and discrete incentives also leads to a very similar prediction of effort (Section 6).

Pay Enough or Don't Pay At All. In the first behavioral treatment we pay a very low piece rate: 1 cent for every 1,000 points. For comparison, the 1-cent benchmark treatment pays 1 cent per 100 points, and thus has *ten* times higher incentives. We examine whether this very low piece rate crowds out motivation as in Gneezy and Rustichini (2000). Crowd-out is captured by a negative Δs_{CO} in expression (4).

To estimate the extent of crowd-out, we predict the counterfactual effort given the incentive, assuming no crowd-out (that is, zero Δs_{CO}): $\hat{e}_{CO} = ((\hat{s} + .001)/\hat{k})^{1/\hat{\gamma}}$.[21] Figure 2b displays the counterfactual predicted effort, 1,893, at the intersection of the marginal cost curve with the marginal benefit set at $\hat{s} + .001$. The model with exponential cost of effort makes a very similar prediction (Appendix Figure 3b), as do models allowing for heterogeneity and discrete incentives (see Section 6 and Appendix A).

[20] The standard errors for the parameters are derived via a bootstrap with 1,000 draws.

[21] As piece rate we use one tenth the piece rate for the benchmark one-cent treatment ($p = .01$), ignoring the fact that the piece rate paid only every 1,000 points. We return to this later in Appendix A.

Remarkably, the observed effort, 1,883, equals almost exactly the counterfactual effort due to incentives. We interpret this evidence as suggesting that crowd-out of motivation in this setting is small, if any. For sure, there is no evidence of large enough crowd-out to reduce effort relative to no piece rate, as in Gneezy and Rustichini (2000): the observed effort in this condition (1,883) is well above the effort with no piece rate (1,521).

Social Preferences. Next, we consider the two charitable giving treatments, in which the Red Cross receives 1 cent (or 10 cents) per 100 points. Figure 3 shows the average effort for all 18 treatments, ranked by average effort. The 1-cent charity treatment induces effort of 1,907, well above the no-piece rate benchmark, but below the treatment with a private 1-cent piece rate. This indicates social preferences with a smaller weight on a charity than on oneself.

Interestingly, the 10-cent charity treatment induces almost identical effort, 1,918. Thus, the social preferences are much more in line with a warm-glow model à la Andreoni (1989; 1990) than with a pure altruism model. Imas (2014) and Tonin and Vlassopoulos (2015) similarly find that, when subjects are earning for a charity, there is no response to the charity return.

The final social preference manipulation is a gift exchange treatment: subjects receive an unexpected bonus of 40 cents, *unconditional* on performance. As Figure 3 and Table 4 show, this treatment, while increasing output relative to the no-pay treatment, has the second smallest effect, 1,602, after the benchmark no-piece-rate treatment. This evidence is consistent with the recent findings suggesting a limited role for reciprocity with positive monetary gifts.

Time Preferences. The two time preference treatments mirror the 1-cent benchmark treatment, except that the promised amount is paid in two (or four) weeks. Figure 3 shows that the temporal delay in the payment lowers effort somewhat, but the effect is quantitatively quite small. More importantly, we do not appear to find evidence for a beta-delta pattern: if anything, the decline in output is larger going from the two-week treatment to the four-week treatment than from the immediate pay to the two-week payment.

Reference Dependence. Next, we focus on loss aversion with treatments that vary the framing of a bonus at a 2,000 threshold as a gain or loss. As Figure 3 shows, the effort is higher for the 40-cent loss framing than for the 40-cent gain framing, replicating the findings of Hossain and List (2012), though the difference is small and not statistically significant. We compare this difference with the effect of increasing the incentives from 40 to 80 cents within the gain framing. In terms of induced output, the 40-cent loss treatment is about halfway between the 40-cent gain treatment and the 80-cent gain treatment. We return in the next section to the implied loss aversion coefficient.

The next set of treatments targets probability weighting, another key component of prospect theory. The probability weighting function magnifies small probabilities but does little to alter probabilities around 0.5. As such, we designed two treatments with stochastic piece rates yielding (in expected value) the same incentive as the 1-cent benchmark, but varying in the probability: a treatment with a 1 percent probability of a $1 piece rate (per 100 points) and

another with a 50 percent probability of a 2 cent piece rate (also per 100 points). Under probability weighting (and approximate risk neutrality), the 1-percent treatment should have the largest effect, even compared to the 1-cent benchmark.

We find no support for this prediction: the treatment with 1 percent probability of $1 yields significantly lower effort (1896) compared to the benchmark 1-cent treatment (2029) or the 50-percent treatment (1,977). Thus, the data does not provide evidence of overweighting of small probabilities; we return to these treatments below.

Psychology-based Treatments. Lastly, we turn to the more psychology-motivated treatments, which offer purely non-monetary encouragements: social comparisons (Cialdini et al., 2007), ranking with other participants, and emphasis of task significance (Grant, 2008).

All three treatments outperform the standard treatment with no piece rate by 200 to 300 points. They also are more effective than the (equally unincentivized) gift-exchange treatment. At the same time, they are less effective than any of the treatments with incentives, including even the very-low-pay treatment. At least in this particular task with MTurk workers, purely psychological interventions have only a moderate effectiveness relative to the power of incentives. Still, they are cost-effective as they increase output for no additional cost.

Comparing across the three treatments, the most effective is the Cialdini-based social comparison treatment which achieves 1,848 points, ahead of the other treatments that reach about 1,750 points. Social comparison treatments have indeed been very popular nudges (Thaler and Sunstein, 2009), used for example by OPower for energy conservation (Allcott, 2008).

Bayesian Shrinkage. One concern with our estimates is that some of the variation in average effort across treatments is due to sampling error. Given our large sample size and the previously reported standard errors, this sampling error is likely to be small. Nonetheless, we can quantify the magnitude by performing a Bayesian shrinkage correction (e.g. Jacob and Lefgren, 2008). Specifically, for each treatment k = 1, ..., 18 we calculate:

$$\hat{e}_{Shrink}=\frac{\bar{\sigma}^2}{\bar{\sigma}^2+\sigma_k^2}\,\hat{e}_k+\left(1-\frac{\bar{\sigma}^2}{\bar{\sigma}^2+\sigma_k^2}\right)\bar{e},$$

where $\bar{\sigma}^2$ is the variance across the 18 effort estimates ($\hat{e}_k$) and σ_k^2 is the square of the estimated standard error of effort for treatment k. The estimator takes a convex combination between the estimated $\hat{e}_k$ (Table 4) and the average effort across all 18 treatments ($\bar{e}$). As Appendix Figure 4 shows, this correction barely affects the point estimates, given that the standard errors for each treatment are small relative to the cross-treatment differences.

4.2 Heterogeneity and Timing of Effort

Distribution of Effort. Beyond the average effort, which is the variable that the experts forecast, it is useful to consider the distribution of effort, especially for treatments with discontinuous incentives. Appendix Figure 5 shows a histogram of effort across all 18 treatments.

Relatively few workers do fewer than 500 presses, and even fewer score more than 3,000 points with almost no one above 3,500 points. There are spikes at each 100 and especially at each 1,000-point mark, in part because of discrete incentives at these round numbers.

Figure 4a presents the cumulative distribution function for the benchmark treatments and for the crowd-out treatment.[22] Incentives induce a clear rightward shift in effort relative to the no-pay benchmark, even with the very low 1-cent-per-1,000-points piece rate. The piece rates are particularly effective at reducing the incidence of effort below 1,000 points, from 20 percent in the no-pay benchmark to less than 8 percent in any of the piece rate conditions.

Figure 4b shows that the treatments with no monetary incentives shift effort to the right, though not as much as the piece rate treatments do. Despite the absence of monetary incentives, there is some evidence of bunching at round numbers of points.

Figure 4c displays the distribution for the gain-loss treatments. We observe, as expected, bunching at 2,000 points, the threshold level for earning the bonus, and missing mass to the left of 2,000 points. Compared to the 40-cent gain treatment, both the 80-cent gain and the 40-cent loss treatments have 5 percent less mass to the left of 2,000 points, and more mass at 2,000 points (the predicted bunching) and points in the low 2,000s. The difference between the three treatments is smaller for low effort (below 1,500 points) or for high effort (above 2,500 points).[23] This conforms to the model predictions: individuals who are not going to come close to 2,000 points, or individuals who due to intrinsic motivation were planning to work hard nonetheless, are largely unaffected by the incentive change. These findings are in line with evidence on bunching and shifts due to discrete incentives and loss aversion (e.g., Rees-Jones, 2014 and Allen, Dechow, Pope, and Wu, forthcoming).[24]

Effort Over Time. As final piece of evidence on the MTurker effort, in Figures 5a and 5b we display the evolution of effort over the 10 minutes of the task. Overall, the average effort remains relatively constant, potentially reflecting a combination of fatigue and learning by doing. The only treatments that, not surprisingly, experience a substantial decrease of effort in the last 3 minutes are the gain/loss treatments, since the workers are likely to have reached the 2,000 threshold by then. The plots also show a remarkable stability in the ranking of the treatments over the different minutes: for example, at any given minute, the piece rate treatments induce a higher effort than the treatments with non-monetary pay. The one

[22]The c.d.f. of effort for the 4-cent treatment, which would be hard to see in the figure, lies between the 1-cent and the 10-cent benchmarks.

[23]Formally, there should be no impact of the change in incentive on the distribution of points about 2,000. However, some small slippage from the threshold at 2,000 is natural.

[24]A comparison with the no-piece rate benchmark also shows that the threshold incentive doubles the share of workers exerting effort above 2,500 points. This difference is not predicted by a simple reference-dependence model, given that there is no incentive to exert effort past the 2,000-point threshold. For the estimation of reference dependence, we compare the three threshold treatments to each other and thus do not take a stand on the level of effort induced by the threshold itself.

exception is the crowd-out treatment which in the final minutes declines in effectiveness.

5 Expert Forecasts

5.1 Mean Expert Forecasts

Which of these results did the experts anticipate? What are the biggest discrepancies? For each treatment, Figure 6 and Table 4 indicate the mean forecast across the 208 experts, along with the actual effort. Table 4 also displays whether there is a statistically significance difference between the mean forecast and the effort.

The largest discrepancy (more than 200 points) between mean forecast and effort is for the crowd-out treatment: on average, experts expect crowd out with a very low piece rate, at least with respect to the counterfactual computed above. Instead, we find no evidence of crowd out.

The next largest deviations occur for the gain-loss treatments: experts expect these treatments to induce an effort of around 2,000 points while the observed effort is around 2,150 points. Notice that this deviation reflects an incorrect expectation regarding the effect of the threshold, *not* a discrepancy about the gain-loss framing. Regarding the latter, the forecasters on average expect about the same effort from the 80-cent gain treatment (2,007) and from the 40-cent loss treatment (2,002). We return to this in the next Section.

The other sizeable deviation is for the gift exchange treatment which, as we noted, has a very limited effect on productivity. Forecasters on average expect an impact of gift exchange that is 107 points larger, 1,709 points versus 1,602 points.

That being said, the experts are remarkably accurate in their forecasts of the psychology-inspired treatments with no incentives: social comparison, ranking, and task significance.

Turning to the charitable giving treatments, the experts are spot on (on average) with their forecast for the 1-cent charitable giving treatment, 1,894 versus 1,907 points. They however predict that the 10-cent charitable giving treatment will yield output that is about 80 points higher, whereas the output is essentially the same under the two conditions. The forecasters expect pure altruism to play a role, while the evidence points almost exclusively to warm glow. We decompose formally the two components in the next section.

It is interesting to consider together all the six treatments with no private monetary incentives: gift exchange, the psychology-based treatments, and the charitable-giving treatments. The experts are remarkably accurate: the average forecast ranks the six treatments in the exact correct order of effectiveness, from gift exchange (least effective) to 10-cent charitable giving (most effective). Furthermore, the deviation between average forecast and actual performance is at most 107 points, a deviation of less than 7 percent from the actual effort.

Considering then the time preference treatments, the experts expect a significant decrease in output with a 2-week delay, compared to the benchmark treatment with the same 1-cent

incentive but no delay. They expect then a small effect (comparatively) of the further 4-week delay. The experts thus anticipate present bias driving a wedge between immediate versus future payments, not between future payments. Instead, the evidence is more consistent with delta discounting. We return to this in the next section.

The final group of treatments regards probability weighting. The experts on average guess just right the output for the treatment with a 50 percent probability of a 2-cent piece rate per 100 points (1,941 versus 1,977). That is, they expect risk aversion to lower effort somewhat relative to the benchmark treatment with 1 cent piece rate, just as it happens. However, the experts on average expect that the effort will be somewhat higher for the treatment with a 1 percent chance of a $1 piece rate, in the direction predicted by probability weighting (though with a modest magnitude). The evidence, instead, does not support the overweighting of small probabilities predicted by probability weighting.

5.2 Heterogeneity of Expert Forecasts

Heterogeneity by Treatment. We now consider the dispersion of forecasts across experts in Figures 7a-d. This heterogeneity is of interest since it captures the degree of disagreement among experts. For each treatment, the graphs also display the observed average effort (the red circle) and the results for the three benchmarks (the vertical lines).

Two piece rate treatments are polar opposites in terms of expert disagreement (Figure 7a). The 4-cent treatment has the least heterogeneity in forecasts, with a standard deviation across experts of 120 points (Table 4). This is not surprising since one can form a forecast using a straightforward model.[25] At the opposite, the 1-cent-per-1,000-point treatment has the most heterogeneity, with a standard deviation of 262 points. About 35 percent of experts expects strong enough motivational crowd out to yield lower output relative to the no-pay treatment (the first vertical line), while other experts expect no crowd out.

Figure 7a also displays the forecasts for the charity treatments, showing a fair degree of disagreement on the expected effectiveness: 20 percent of experts expects the 1-cent charity treatment to outperform the 1-cent piece rate treatment. These experts expect that workers assign a higher weight on the return to a charity than on an equal-size private return.

Figure 7b presents the evidence for the delayed-payment treatments: nearly all experts think that delayed payment will lower effort, or have no effect.

The results for the probability weighting treatments (also in Figure 7b) reveal substantial heterogeneity. Fifty percent of experts expect higher effort in the 1 percent treatment than in the 1-cent benchmark; of these experts, almost half expects strong enough overweighting of small probabilities to lead to higher effort than in the 10-cent benchmark. The remaining

[25] Nonetheless, about 20 percent of experts guess an output that is lower than the output for the 1-cent treatment or higher than the output for the 10-cent treatment.

fifty percent of experts instead expects risk aversion (over small stakes) to be a stronger force. There is much less variance among experts for the 50-percent treatment, as one would expect, since probability weighting, to a first approximation, should not play a role.

Figure 7c presents the evidence for the gain and loss treatments, showing that the c.d.f.s for the 80-cent gain and the 40-cent loss treatment are right on top of each other.

For the remaining treatments with no incentive pay—gift exchange and the psychology treatments—, there is a fairly wide distribution of guesses mostly between the no-pay treatment and the 1-cent piece rate treatment (Figure 7d). For the two social comparison treatments, in fact, 25 percent of experts expect that these treatments would outperform the 1-cent piece rate treatment. In reality, the treatments, while effective, are not that powerful.

Why do the crowd out and probability weighting treatments have the largest dispersion of opinions? The wide priors for these treatments may be related to the more limited evidence in the literature. Only a small number of papers followed the Gneezy and Rustichini (2000) design. Overweighting of small probabilities, while part of a larger literature, also has not attracted the same attention as, say, present bias or loss aversion.

The dispersion of forecasts in these treatments may also reflect behavioral forces affecting effort in opposite directions, such as overweighting of small probabilities versus curvature of the utility function in the probabilistic pay treatment. If that is the case and the contrasting behavioral forces differ across workers, treatments with high heterogeneity in forecasts may also have high heterogeneity in MTurker effort. Figure 8 shows that this is indeed the case: among the 15 treatments, treatments with a higher standard deviation of MTurker effort also have a higher standard deviation of expert forecasts.

Field. Is the heterogeneity in forecasts explained in part by differences in the field of expertise? Figure 9 presents the average forecast by treatment separately for experts with primary field in behavioral economics, laboratory experiments, standard economics, and psychology and decision-making. Perhaps surprisingly, the differences are small. All groups of experts expect more crowd out than in the data, expect more gift exchange than in the data, and expect higher effort for the 10-cent charitable giving treatment compared to the 1-cent charitable giving treatment. There are some differences—psych experts expect less overweighting of small probabilities—, but the differences are small and unsystematic. Field of expertise, thus, does not explain the heterogeneity in forecasts.[26]

6 Estimates of Behavioral Parameters

An advantage of field experiments is that their design can be tightly tailored to a model, so as to test the model and estimate parameters. Surprisingly, such model-based field experiments are

[26] In DellaVigna and Pope (2016) we consider further characteristics, such as citations and academic rank.

still relatively uncommon (Card, DellaVigna, and Malmendier, 2011). One of the difficulties of conducting these field experiments is that the researcher needs to be able to estimate a set of nuisance parameters (e.g., about the environment or ancillary preferences), in order to focus on the parameters of interest. Creating the appropriate variation is not always straightforward.

In our experiment, the simplicity of the chosen task implies that the only nuisance parameters which the researcher needs to control for are those on the cost of effort, and the baseline motivation. We thus designed the piece rate treatments to pin down these parameters, as stressed in Section 4. Armed with these estimates, we can then use the model to cast quantitative light on the behavioral parameters of interest. Furthermore, since we informed the experts about the results in the benchmark treatments, we can, at least in principle, assume that the forecasters approximately share the model and the estimates for these nuisance parameters.

The treatments provide evidence on three key models in behavioral economics: the beta-delta model of time preferences (Laibson, 1997; O'Donoghue and Rabin, 1999), the reference-dependent model of risk preferences (Kahneman and Tversky, 1979), and altruism and warm glow models of social preferences (Becker, 1974; Andreoni, 1989, 1990). We revisit the qualitative evidence on these models documented in Sections 4 and 5, and back out key parameters.

We should stress that some of the treatments are not ideal tests for the behavioral models, given the constraints of the setting. In particular, the estimates for present bias should be obtained from trade-offs of utility and consumption (as in Augenblick, Niederle, and Sprenger, 2015), not from trade-offs of monetary payments. Having said that, our setting has two unique advantages. First, we are able to back out key parameters for a range of behavioral models in a single setting; in contrast, the papers that estimate behavioral parameters typically have evidence on just one behavioral model. Second, we compare estimates for the observed behavior of MTurk workers to the estimates implied by the expert forecasts, a unique feature.

We employ two estimation procedures. In the first one, mentioned in Section 4, we use a minimum distance estimator with the average points by treatment as moments. The advantage of this procedure is that the experts, in principle, could also estimate this model, since they observe the average effort for the benchmark treatments before making the forecasts.

A disadvantage of this procedure is that it assumes, counterfactually, no heterogeneity in effort within a treatment. It also assumes, for simplicity, that the incentives accrue continuously, as opposed to at fixed 100-point intervals. In a second procedure, we allow for heterogeneity in the marginal cost of effort and estimate the model on individual effort data using non-linear least squares. We now present the two estimation procedures, and the resulting estimates.

Minimum-Distance Estimation. For the minimum-distance estimation, we use as moments the average effort in the three benchmark treatments (no-pay, 1-cent, and 10-cent) to estimate $\hat{\gamma}$, $\hat{s}$, and $\hat{k}$. We estimate the model under the assumption of power cost of effort function and under the assumption of exponential cost of effort. Panel A of Table 5 presents the estimates in Columns 1 and 3, as we discussed in Section 4.

Given these estimates, in a second stage we back out the behavioral parameters using the average effort in the relevant treatments as moments. Consider for example the altruism and warm glow parameters, α and a, for the power cost of effort case. Effort in the 1-cent and 10-cent charitable giving treatments equal

$$\bar{e}_{CH.01} = \left(\frac{\hat{s} + (\hat{a} + \hat{\alpha}) * .01}{\hat{k}}\right)^{1/\hat{\gamma}} \text{ and } \bar{e}_{CH.10} = \left(\frac{\hat{s} + \hat{a} * .01 + \hat{\alpha} * .10}{\hat{k}}\right)^{1/\hat{\gamma}}. \tag{11}$$

The system of two equations in two unknowns (given the estimates of $\hat{\gamma}$, $\hat{s}$, and $\hat{k}$) yields solutions for $\hat{\alpha}$ and $\hat{a}$. By design, the model is just identified.[27] We derive confidence intervals for the parameters using a bootstrap procedure.[28]

The appeal of this simple identification strategy is that the forecasters could also, at least in principle, have obtained the same estimates for $\hat{\gamma}$, $\hat{s}$, and $\hat{k}$, given the observed effort in the benchmark treatments. Under this assumption, we can take the forecasts $(e^i_{CH.01}, e^i_{CH.10})$ of expert i and back out the implied beliefs about social preferences $(\widetilde{\alpha}_i, \widetilde{a}_i)$ of expert i.

Non-Linear Least Squares. As we discussed above, the minimum-distance estimate assumes no error term, smooths incentives over the continuum of points, instead of modelling discrete incentives at the 100-point thresholds, and only uses mean effort levels rather than the individual effort provision. We now relax these assumptions.

We allow for a heterogeneous marginal cost of effort $c(e)$ in maximization problem (1). Namely, for the power cost case we assume that $c(e) = ke^{1+\gamma}(1+\gamma)^{-1}\exp(-\gamma\varepsilon)$, with ε normally distributed $\epsilon \sim N(0, \sigma^2_\epsilon)$. The additional noise term $\exp(-\gamma\epsilon)$ has a lognormal distribution, ensuring positive realizations for the marginal cost of effort. As DellaVigna, List, Malmendier, and Rao (2015) show, this implies the first-order condition $s+p-ke^\gamma\exp(-\gamma\epsilon) = 0$ and, after taking logs and transforming,

$$\log(e) = \frac{1}{\gamma}[\log(s+p) - \log(k)] + \epsilon. \tag{12}$$

Equation (12) can be estimated with non-linear least squares (NLS). Similarly, for the case of exponential cost function we assume $c(e) = k\exp(\gamma e)\gamma^{-1}\exp(-\gamma\varepsilon)$, yielding a parallel estimating expression but with effort, rather than log effort, as dependent variable:

$$e = \frac{1}{\gamma}[\log(s+p) - \log(k)] + \epsilon. \tag{13}$$

[27] Jointly estimating a system of five equations in five unknowns including also the three benchmark treatments yields identical estimates.

[28] We draw 1,000 samples. In each bootstrap iteration we resample (with replacement) workers from each treatment, form a new sample with the same number of observations as the original, and calculate the mean effort. We then re-estimate the parameters k, γ, and s, as well as the relevant behavioral parameters. For the confidence intervals we use the 2.5th percentile and the 97.5th percentile of the estimated parameters across the 1,000 iterations.

The NLS estimation allows us to model the heterogeneity in effort. To take into account the discontinuous incentives, we assume that the individual chooses output in units of 100 points, and estimate the model using output rounded to the closest 100-point: that is, a score of 2,130 points is recorded as 21 units of 100 points.[29] This assumption allows us to use the first-order condition for effort and thus the non-linear least squares for estimation.[30]

Columns 2 and 4 of Panel A in Table 5 display the estimates of the non-linear least squares model using the benchmark treatments. The parameter estimates for the exponential cost function case (Column 4) are nearly identical to the minimum-distance ones (Column 3). The model perfectly fits the benchmark treatments and makes predictions for the 4-cent treatment and for the low-pay treatment that are very similar to the minimum distance ones.[31]

The estimates for the power cost function (Column 2) are somewhat different from the minimum-distance one, as one would expect given Jensen's inequality. The NLS model, as (12) stresses, matches the expected log effort, while the minimum-distance matches the log of expected effort (given the assumed homogeneity). Taking into account heterogeneity in the NLS estimates leads to a somewhat lower curvature ($\hat{\gamma}_{NLS} = 24$ versus $\hat{\gamma}_{MD} = 33$). Nonetheless, both models fit the in-sample moments perfectly and make similar predictions for the 4-cent treatment and for the low-pay treatment.[32]

We use the NLS estimator to estimate the behavioral parameters in Panel B. Formally, we run a NLS regression including the benchmark treatments as well as the behavioral treatments. We report the point estimates for the behavioral coefficients (Columns 3 and 6) and, for the exponential case, the behavioral parameters implied by the expert forecasts (Column 7).[33]

Social Preferences. Returning to the social preference example, equations (11) clarify the difference between altruism and warm glow: the altruism parameter α multiplies the actual return to the charity while the warm glow term a multiplies a constant return which we set, for convenience of interpretation, equal to .01, the 1-cent return. Taking logs of output in both

[29]Scores in the 50-99 range are rounded up, while scores in the 0-49 range are rounded down. For the very first bin, that is, scores of 1-49 points, we round to the midpoint, 25.

[30]Notice that this estimation strategy, while not making use of the full information, is not mis-specified, as it recognizes incentives as actually set. The alternative is to model the continuous point score using maximum-likelihood, model the bunching at the 100-point score. We opted for the simpler and more transparent non-linear least squares estimate.

[31]The implied effort for the low-pay treatment still assumes an incentive of .1 cent every 100 point, rather than an incentive ocurring only every 1,000 points. In Appendix A we show that modelling the discrete jumps at 1,000 gives similar results for the implied effort in the low-pay treament.

[32]Notice that for the NLS model with power cost in Column 2 of Table 5, the predictions are evaluated using the average log effort.

[33]For the power cost case we cannot infer the parameters implied by the expert forecasts since we did not elicit the expected log points, as the model requires.

treatments and differentiating, we obtain

$$\log\left(\bar{e}_{CH.10}\right) - \log\left(\bar{e}_{CH.01}\right) = \frac{1}{\hat{\gamma}}\left[\log\left(\hat{s} + \hat{a} * .01 + \hat{\alpha} * .10\right) - \log\left(\hat{s} + (\hat{a} + \hat{\alpha}) * .01\right)\right].$$

Thus, the percent increase in output between the two treatments (left-hand side) identifies the altruism parameter α, since the two log terms in the right-hand side differ only in the terms $\hat{\alpha} * .10$ versus $\hat{\alpha} * .01$. The warm glow parameter $\hat{a}$ is identified from the level of effort in the 1-cent charity treatment. The expression also makes clear that $1/\hat{\gamma}$ is the elasticity of effort with respect to motivation.

Panel B of Table 5 shows that the altruism coefficient from the MTurk effort is estimated to be essentially zero in all four specification, e.g. $\hat{\alpha} = 0.003$ in Column 1. Importantly, the confidence interval is tight enough that we can reject even small values, such as the workers putting .03 as much weight on the charity as on themselves (Column 1). The distribution of beliefs among experts is quite different: the median expert expects altruism $\widetilde{\alpha}_{med} = 0.067$ (Columns 2 and 5), outside the confidence interval of the MTurk estimates.

The pattern for warm glow is the converse: the worker effort indicates sizable warm glow, with a weight $\hat{a}$ between 0.12 (Column 1) and 0.20 (Column 3) on the average return for the charity. The median forecast instead is $\widetilde{a}_{med} = 0.02$ (Column 1), which is barely inside the 95 percent confidence interval for the estimates from the MTurk effort. The median expert expects a pure altruism model, counterfactually.

Figures 10a-b show the distribution of the social preferences parameters $(\widetilde{\alpha}_i, \widetilde{a}_i)$ estimated from the 208 expert forecasts from the minimum-distance power cost specification (Column 1). The green solid line denotes the value implied by the median forecast, and the red dashed line indicates the parameter value implied by the actual MTurk worker effort.

The discrepancy between the experts and the data is interesting because the findings from MTurk effort are consistent with a long-standing research line suggesting that pure altruism does not capture well charitable giving. It appears that the experts were not fully expecting a conclusion that is consistent with the previous literature.

Panel B of Table 5 also reports the estimated shift in motivation due to gift exchange. The impact on motivation is estimated to be tiny, consistent with the small gift exchange effect, as well as the small value for baseline motivation. We do not report the other motivation shift parameters in response to the other non-monetary treatments, but the estimates are similarly small in magnitude. The expert forecasts are generally in line, though some experts expect a sizeable shift in motivation due to the treatments.

Time Preferences. To estimate the present-bias parameters, we model effort in the delayed-payment treatments as in (7), with t denoting the weeks of delay, β the present bias parameter, and δ the (weekly) discount factor. As Panel B of Table 5 indicates, the estimates of the time preference parameters from the worker effort are noisy: the point estimate indicates no present bias, but the confidence intervals for β are wide. Even given the imprecise estimate

from the MTurk data, there is useful information in the expert forecasts: as Figures 10c-d show for the specification in Column 2, the median expert expects present bias ($\widetilde{\beta}_{med} = 0.76$) with a significant left tail of smaller estimates (as well as estimates above 1). These expectations appear anchored on earlier experimental results on time preferences, which differ from the more recent results of no present bias with respect to monetary payments (Andreoni and Sprenger (2012); Augenblick, Niederle and Sprenger (2015).

Probability Weighting. In prospect theory, the probability weighting function $\pi(P)$ transforms probabilities P into weights, which are then used to calculate the value of the 'prospects'. The evidence on probability weighting (e.g., Prelec, 1998) suggests that small probabilities are overweighted by a factor of 2 to 4, with probabilities around 50 percent left unaltered. The treatment with a 1 percent probability of a \$1 piece rate allows us to test for such overweighting of small probability and estimate $\pi(0.01)$. The design also includes a treatment with 50 percent probability of a 2-cent piece rate to provide evidence on the concavity of the value function, i.e., the risk aversion.

We model optimal effort in the probabilistic treatments as in (10), allowing for a possibly concave utility function $u(p) = p^{\theta}$. This includes linear utility ($\theta = 1$), assumed so far, as well as the calibrated value $\theta = 0.88$ from Tversky and Kahneman (1991). We assume that the probability weight does not transform the 50-percent probability ($\pi(0.5) = 0.5$).

Since allowing for curvature in the utility function $u(p)$ affects the estimates also in the benchmark treatments, we re-estimate the baseline parameters as well as the probability weighting parameters using observations in the three benchmark treatments and in the two probabilistic treatments. In Table 6, Panel A we report the results for the non-linear least squares estimates; the results are similar with the minimum-distance estimates.

The probability weight for a 1 percent probability is estimated to be *smaller* than 1 percent under the assumption of either linear utility (Columns 1 and 4) or concave utility with the Kahneman and Tversky curvature (Columns 2 and 5). Thus, we do not find evidence of overweighting of small probabilities. In contrast, the median expert expects overweighting of 1 percent probability under either specification (Columns 4 and 5). The difference between the median forecast and the estimate from the MTurk effort is statistically significant.

The specification with estimated curvature of the utility function (Columns 3 and 6) leads to imprecise results, yielding (implausibly) high curvature with the exponential cost function (Column 6) and near-linear utility with power cost function (Column 3). The former case, given the high curvature of the value function, is the only case with estimates implying overweighting of small probability, but the estimates are very imprecise. Figure 10e displays the distribution of the implied probability weights under the three assumptions.

To summarize, under plausible values for the curvature of the value function, the MTurk effort does not provide evidence of overweighting of small probabilities, contrary to the expectations of the median expert.

Loss Aversion. We estimate the loss aversion parameter λ using the three gain-loss treatments. As Figures 6 and 7c show, the experts are quite off in their forecasts of these treatments because it was difficult to predict the impact of a threshold payment at 2,000 points.[34] For our estimation strategy, we employ an approximation that bypasses the misprediction of the effect of the threshold payment. We identify the loss aversion comparing two responses: the difference between the 40-cent loss treatment and the 40-cent gain treatment $e_{L.40} - e_{G.40}$, and the difference between the 80-cent gain treatment and the 40-cent gain treatment, $e_{G.80} - e_{G.40}$. As we show in Appendix A, the following approximation holds

$$\frac{e_{L.40} - e_{G.40}}{e_{G.80} - e_{G.40}} \simeq \frac{(\lambda - 1)\,\eta}{1 + \eta}.$$

Under the standard assumption of unitary gain utility ($\eta = 1$), this expression allows for estimation of the loss aversion λ.[35]

The distribution of the loss aversion parameter $\widetilde{\lambda}_i$ according to the experts is broadly centered around 2.5-3, with a median $\widetilde{\lambda}_{med} = 2.75$ (Figure 10f and Table 6, Panel B). Thus, experts hold beliefs in line with the Tversky and Kahneman (1991) calibration which, revisited in the Koszegi and Rabin (2006) formulation, implies a loss aversion parameter of $\lambda = 3$ (assuming $\eta = 1$). The estimate from the MTurk worker effort is smaller, $\hat{\lambda} = 1.73$, but with a wide confidence interval including the value $\lambda = 3$. Unfortunately, the estimate for λ is quite noisy because the impact of going from the 40 cent gain treatment to the 80 cent gain treatment is quite small, making it hard to compare to the effect of the 40 cents loss treatment.

Robustness. In Appendix Table 1 we explore the robustness of the estimated parameter values to alternative specifications. We present the results for the non-linear least squares specification with exponential cost of effort function; the results are parallel with the other specifications. First, we examine the impact of mis-specification in the cost function by forcing the curvature parameter γ to the values of .01 (Column 1) and .02 (Column 2). Second, we allow for curvature of the value function with concavity $\theta = 0.88$ when estimating the parameters (Column 3). Third, we use continuous points assuming that the piece rates are paid continuously (Column 4). These changes have limited impact on the social preference estimates. Furthermore, the implied effort under the low-pay treatment implies no crowd-out of motivation, as in the benchmark estimates. The estimated present-bias coefficient β is more sensitive, not surprisingly given the wide confidence intervals in the benchmark estimates.

[34] In hindsight, we should have offered the results of the 40 cent gain treatment as a fourth benchmark.

[35] Unlike the other derivations, this solution is an approximation. However, given that the differences in effort between the threshold treatments are small, the bias in estimate due to the approximation should be small as well. Given that the estimation is based on a ratio, we only use observations in which the denominator is positive and larger than 10 units of effort, since smaller differences may be hard for experts to even control with a mouse. We also do not include observations with negative λ.

7 Conclusion

What motivates workers in effortful tasks? How do different monetary and non-monetary motivators compare in effectiveness? Do the results line up with the expectations of researchers?

We present the results of a large-scale real-effort experiment on MTurk workers. The model-based 18-arm experiment compares three classes of motivators: (i) standard incentives in the form of piece rates; (ii) behavioral factors like present-bias, reference dependence, and social preferences, and (iii) non-monetary inducements more directly borrowed from psychology.

We find that monetary incentives work as expected, including a very low piece rate treatment which does not crowd out motivation. The evidence is generally consistent with standard behavioral models, including loss aversion and warm glow, though we do not find evidence of overweighting of small probabilities. The psychological motivators are effective, though less so than monetary incentives.

We then compare the results to forecasts by 208 behavioral experts. The experts on average anticipate several key features of the data, like the effectiveness of psychological motivators compared to the effectiveness of incentives. A sizeable share of the experts, however, expect crowd-out, probability weighting, and pure altruism, unlike what we observe in the data.

An important caveat is that the relative effectiveness of the various treatments is likely to be context dependent. Some treatments that had a limited effect on motivation in our context, such as probability weighting, may have large effects in a different task or with a different participant pool (non-Mturk workers). As always, it will be important to see replications. By estimating the structural behavioral parameters, we set up a methodology that should allow the comparison of effects across different settings and subject pools.

Further, while we have studied a large set of behavioral motivators in this paper, it is by no means an exhaustive list. For example, we considered but ultimately did not include treatments related to limited attention and salience, left-digit bias, and overconfidence among others. In addition, our focus has been on costly effort, but one could similarly consider the impact of incentives and behavioral motivators on other outcomes, like contributions to public goods. Future work can hopefully extend the setting in this paper in some of these directions.

Finally, the combination of head-to-head comparisons of treatments and expert forecasts can help inform the role of behavioral economists in helping policy-makers or businesses. For example, one of the authors recently worked with a non-profit company that was trying to motivate its clients to refinance their homes during a period of low interest rates. The company wanted advice on the design of a letter to send to their clients in order to maximize take up. Should we trust our intuition and recommend a phrasing for the letter, or should we strongly recommend that the company run an RCT to test possible options? Of course, when possible a randomized trial in the relevant setting is ideal, but studies such as ours start to suggest the extent to which researcher expectations of future results are accurate.

References

[1] Allcott, Hunt. 2012. "Social Norms and Energy Conservation." *Journal of Public Economics*, Volume 95, Issues 9–10, October 2011, Pages 1082–1095.

[2] Allen, Eric J., Patricia M. Dechow, Devin G. Pope, George Wu. Forthcoming. "Reference-dependent preferences: Evidence from marathon runners," *Management Science.*

[3] Amir, On, and Dan Ariely. "Resting on Laurels: The Effects of Discrete Progress Markers as Subgoals on Task Performance and Preferences." *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Vol. 34(5) (2008), 1158-1171.

[4] Andreoni, James. 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy*, 97(6), 1447-1458.

[5] Andreoni, James. 1990. Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal*, 100(401), 464-477.

[6] Andreoni, James and Charles Sprenger. 2012. "Estimating Time Preferences from Convex Budgets," *American Economic Review*, vol. 102(7), pages 3333-56.

[7] Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. "No Margin, No Mission? A Field Experiment on Incentives for Pro-Social Tasks." *Journal of Public Economics* Vol. 120 (2014): 1-17.

[8] Augenblick, Ned, Muriel Niederle and Charles Sprenger. 2015. "Working Over Time: Dynamic Inconsistency in Real Effort Tasks " *Quarterly Journal of Economics*, 130 (3): 1067-1115.

[9] Bandiera, Oriana, Iwan Barankay, and Imran Rasul. "Team Incentives: Evidence from a Firm Level Experiment." *Journal of the European Economic Association* Vol. 11, No. 5 (2013): 1079-1114.

[10] Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg. 2016. Forthcoming. "Decision Theoretic Approaches to Experiment Design and External Validity", *Handbook of Field Experiments.*

[11] Barberis, Nicholas C. 2013. "Thirty Years of Prospect Theory in Economics: A Review and Assessment." *Journal of Economic Perspectives*, 27(1): 173-96.

[12] Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C. Teitelbaum. 2013. "The nature of risk preferences: Evidence from insurance choices." *American Economic Review* 103, no. 6 (2013): 2499-2529.

[13] Becker, Gary S. 1974. "A Theory of Social Interactions" *Journal of Political Economy,* 82(6), 1063-1093.

[14] Berger, Jonah, and Devin Pope. "Can Losing Lead to Winning." *Management Science* Vol. 57(5) (2011), 817-827.

[15] Camerer, Colin et al.. 2016. "Evaluating Replicability of Laboratory Experiments in Economics" *Science*, 10.1126.

[16] Card, David, Stefano DellaVigna, and Ulrike Malmendier. 2011. "The Role of Theory in Field Experiments". *Journal of Economic Perspectives,* 25(3), pp. 39-62.

[17] Cialdini, Robert M., et al. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science* Vol. 18, No. 5 (2007).

[18] Coffman, Lucas and Paul Niehaus. 2014. "Pathways of Persuasion" Working paper.

[19] Conlin, Michael, Ted O'Donoghue, and Timothy J. Vogelsang. 2007. "Projection Bias in Catalog Orders." *American Economic Review,* 97(4), 1217-1249.

[20] Deci, Edward L. "Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology* Vol. 18, No. 1 (1971): 105-115.

[21] DellaVigna, Stefano. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature* Vol. 47, No. 2 (2009): 315-372.

[22] DellaVigna, Stefano, List, John. A., & Malmendier, Ulrike. (2012). "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics*, 127(1), 1-56.

[23] DellaVigna, Stefano, John List, Ulrike Malmendier, Gautam Rao. 2015. "Estimating Social Preferences and Gift Exchange at Work" Working paper.

[24] DellaVigna, Stefano and Devin Pope. 2016. "Predicting Experimental Results: Who Knows What?" Working paper.

[25] Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. "Using prediction markets to estimate the reproducibility of scientific research", *PNAS,* Vol. 112 no. 50, 15343–15347.

[26] Erev, Ido, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan M. Herzog, Robin Hau, Ralph Hertwig, Terrance Stewart, Robert West, and Christiane Lebiere, "A Choice Prediction Competition: Choices from Experience and from Description." *Journal of Behavioral Decision Making*, 23 (2010): 15-47.

[27] Esteves-Sorenson, Constanca. 2015. "Gift Exchange in the Worplace: Addressing the Conflicting Evidence with a Careful Test " Working paper.

[28] Fehr, Ernst and Simon Gächter, 2000. "Fairness and Retaliation - The Economics of Reciprocity", *Journal of Economic Perspectives*, 14, 159-181.

[29] Fehr, Ernst, Georg Kirchsteiger and Arno Riedl. 1993. "Does Fairness Prevent Market Clearing? An Experimental Investigation", *Quarterly Journal of Economics.* Vol. 108, No. 2, pp. 437-459.

[30] Frank, R.H. 1985. *Choosing the Right Pond: Human Behavior and the Quest for Status.* New York: Oxford University Press.

[31] Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature* Vol. 40, No. 2: 351-401.

[32] Fryer, Roland Jr., Stephen D. Levitt, John A. List, and Sally Sadoff. (2012). "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." NBER Working Paper No. 18237.

[33] Uri Gneezy and John A List. 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments" *Econometrica*, Volume 74, Issue 5, pages 1365–1384.

[34] Gneezy, Uri and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics* Vol. 115, No. 3: 791-810.

[35] Grant, Adam M. 2008. "The Significance of Task Significance: Job Performance Effects, Relational Mechanisms, and Boundary Conditions." *Journal of Applied Psychology* Vol. 93, No. 1: 108-124.

[36] Groh, Matthew, Nandini Krishnan, David McKenzie, Tara Vishwanath. 2015. "The Impact of Soft Skill Training on Female Youth Employment: Evidence from a Randomized Experiment in Jordan" Working paper.

[37] Horton, John J. and Chilton, Lydia B. 2010. "The Labor Economics of Paid Crowdsourcing" *Proceedings of the 11th ACM Conference on Electronic Commerce.*

[38] Hossain, Tanjim and List, John A. 2012. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Science* Vol. 58, No. 12: 2151-2167.

[39] Imas, Alex. 2014. "Working for the "warm glow": On the benefits and limits of prosocial incentives" *Journal of Public Economics*, Vol. 114, pp. 14–18.

[40] Ipeirotis, Panagiotis G. "Analyzing the Amazon Mechanical Turk Marketplace. 2010. " *XRDS: Crossroads, The ACM Magazine for Students* Vol. 17, No. 2: 16-21.

[41] Jacob, Brian and Lefgren, Lars. 2008. "Principals as Agents: Subjective Performance Assessment in Education," *Journal of Labor Economics*, 26(1), 101-136.

[42] Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica* Vol. 47, No.2: 263-292.

[43] Koszegi, Botond. 2014. "Behavioral Contract Theory" *Journal of Economic Literature*, 52(4), pp. 1075-1118.

[44] Koszegi, Botond and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences", *Quarterly Journal of Economics*, Vol. 121, No. 4, pp. 1133-1165.

[45] Kube, Sebastian, Michel André Maréchal and Clemens Puppe. 2013. "Do Wage Cuts Damage Work Morale? Evidence From A Natural Field Experiment", *Journal of the European Economic Association* Volume 11, Issue 4, pages 853–870.

[46] Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015. "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105(4): 1478-1508.

[47] Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* Vol. 112, No. 2: 443-477.

[48] Laibson, David, Andrea Repetto, and Jeremy Tobacman. 2007. "Estimating Discount Functions with Consumption Choices over the Lifecycle," Working paper.

[49] Loewenstein, George, Troyen Brennan, and Kevin G. Volpp. 2007. "Asymmetric Paternalism to Improve Health Behaviors." *Journal of the American Medical Association* Vol. 298, No. 20: 2415-2417.

[50] Maslow, Abraham H. 1943. A Theory of Human Motivation. *Psychological Review*, pp. 370-396.

[51] Mullainathan, Sendhil and Richard H. Thaler. 2001. "Behavioral Economics." In *International Encyclopedia of the Social & Behavioral Sciences*: 1094-1100.

[52] O'Donoghue, Edward and Matthew Rabin. 1999. "Doing It Now or Later". *American Economic Review,* Vol. 89(1), 103-124.

[53] Paolacci, Gabriele. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgement and Decision Making* Vol. 5, No. 5: 411-419.

[54] Paolacci, Gabriele, and Jesse Chandler. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* Vol 23(3), 184-188.

[55] Pennisi, Elizabeth. "A Low Number Wins the GeneSweep Pool." *Science* Vol. 300, No. 5625 (2003): 1484.

[56] Prelec, Drazen. 1998. "The Probability Weighting Function." *Econometrica* Vol. 66, No. 3: 497-527.

[57] Rabin, Matthew. 1998. "Psychology and Economics." *Journal of Economic Literature* Vol. 36, No. 1: 11-46.

[58] Rees-Jones, Alex. 2014. "Loss aversion motivates tax sheltering: Evidence from US tax returns" Working paper.

[59] Ross, Joel, et al. 2010. "Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk." In CHI '10 Extended Abstracts on Human Factors in Computing Systems: 2863-2872.

[60] Sanders, Michael, Freddie Mitchell, and Aisling Ni Chonaire. 2015. "Just Common Sense? How well do experts and lay-people do at predicting the findings of Behavioural Science Experiments" Working paper.

[61] Tetlock, Philip E. 2010. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

[62] Tetlock, Philip E., Dan Gardner. 2015 *Superforecasting: The Art and Science of Prediction*, Crown Publisher.

[63] Thaler, Richard H. and Cass R. Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* New York: Penguin Group.

[64] Tonin, Mirco and Michael Vlassopoulos. 2015. "Corporate Philanthropy and Productivity: Evidence from an Online Real Effort Experiment." *Management Science* 61(8):1795-1811.

[65] Tversky, Amos and Daniel Kahneman. 1992. "Advances in prospect theory: Cumulative representation of uncertainty" *Journal of Risk and Uncertainty*, Volume 5, Issue 4, pp 297-323.

[66] Wu, George and Richard Gonzalez. 1996. "Curvature of the Probability Weighting Function." *Management Science* Vol. 42, No. 12: 1676-1690.

A Appendix A - Estimation Appendix

Loss Aversion. A threshold payment (such as at 2,000 points) induces bunching at the threshold, and a missing mass to the left of the threshold. The extent of bunching and the missing mass will be increasing in the utility gain u of achieving the threshold. Denote with $F(u)$ the share bunching as a function of the utility benefit u which, as derived in Section 2, equals $(1+\eta)G$ in the gain treatment and $(1+\lambda\eta)G$ in the loss treatment. The average effort e in a treatment will be an increasing function ϕ of the bunching, and thus $e = g(u)$, where $g(.) \equiv \phi(F(.))$ is an increasing function. Consider a linear approximation to how the average effort responds to a change in u: $de/du = g'(u^*) * du$. The effort change going from the 40-cent gain condition to the 80-cent gain condition is approximately: $e_{G.80} - e_{G.40} \simeq g'(u^*)[(1+\eta).80 - (1+\eta).40] = g'(u^*)(1+\eta) * .40$. Similarly, the effort change going from the 40-cent gain condition to the 40-cent loss condition is approximated as $e_{L.40} - e_{G.40} \simeq g'(u^*)[(1+\lambda\eta).40 - (1+\eta).40] = g'(u^*)(\lambda-1)\eta * .40$. The ratio of these differences is

$$\frac{e_{L.40} - e_{G.40}}{e_{G.80} - e_{G.40}} \simeq \frac{g'(u^*)(\lambda-1)\eta * .40}{g'(u^*)(1+\eta) * .40} = \frac{(\lambda-1)\eta}{1+\eta}.$$

The term $g'(u^*)$ drops out, leaving a function of just λ and η. Under the standard assumption of unitary gain utility ($\eta = 1$), the ratio of the difference in effort allows for estimation of the loss aversion λ. Notice that, unlike the other derivations, this solution is an approximation. However, given that the differences in effort between the threshold treatments are small, the bias in estimate due to the approximation should be small as well.

Given that the estimation is based on a ratio, we only use observations in which the denominator is positive and larger than 10 units of effort, since smaller differences may be hard for experts to even control with a mouse, and we exclude observations with negative λ.

Low-Pay Treatment. The predicted effort for the low-pay treatment in Table 5 assumes for simplicity that the incentive (1-cent every 1,000 points) is paid continuously, as opposed to only at every 1,000-point threshold. This is true also for the NLS specification, which assumes an incentive of 0.1 cents every 100 points in order to be apply the first order conditions.

We now show that modelling the payoff jumps at the 1,000-point thresholds leads to similar predicted effort for the low-pay condition. Consider the NLS estimate with exponential cost of effort function. (Modelling the threshold effects only makes sense for models with heterogeneity, that is, the NLS models and not the minimum-distance model). Individual i maximizes

$$\max_{e_i} se_i + p(\mathbf{1}_{e_i \geq 1000} + \mathbf{1}_{e_i \geq 2000} + \mathbf{1}_{e_i \geq 3000}) - c_i(e_i)$$

where $c_i(e_i)$ is specific to person i: $c_i(e_i) = k\exp(\gamma e_i)\gamma^{-1}\exp(-\gamma\varepsilon_i) = (k/\gamma)\exp(\gamma(e_i - \varepsilon_i))$. The second term models the threshold compensation p (which equals 1 cent in this case). For expositional simplicity, we assume that exceeding 4,000 points is too costly. Consider the optimal solution without incentives ($p = 0$): $s - k\exp(\gamma e_i)\exp(-\gamma\varepsilon_i) = 0$ or

$$e_i = \frac{1}{\gamma}[\log(s/k)] + \varepsilon_i. \tag{14}$$

Given the estimated $\hat{s}, \hat{\gamma}$, and $\hat{k}$, the realization of ε_i pins down uniquely e_i. Thus, denote with $\varepsilon_i(e_i)$ the error term that leads to the choice of e_i with no incentives. We now show that the solution takes a threshold form, which we characterize first with respect to the threshold at 1,000 points. There is a value ε^{1000} such that any type $\varepsilon < \varepsilon^{1000}$ stays at the effort level chosen with no incentive as in (14). Any type with $\varepsilon > \varepsilon^{1000}$ chooses instead to jump to the threshold effort of 1,000 or stay at his already higher level of effort. Label as $U_0(e_i)$ the utility that type

e_i achieves at the optimum as in (14), substituting for the expression for $\varepsilon(e_i)$ and simplifying, we obtain

$$U_0(e_i) = se_i - \frac{k}{\gamma}\exp(\gamma e_i)\exp\left(-\gamma e_i + \log\left(\frac{s}{k}\right)\right) = s\left(\bar{e} - \frac{1}{\gamma}\right).$$

Label as U_{1000} the utility that type e_i achieves from exerting effort 1,000. With similar substitutions and simplifications we obtain

$$U_{1000}(e_i) = 1000s + 1 - \frac{s}{\gamma}\exp(\gamma(1000 - \bar{e})).$$

We show now that there exists one and only one $\bar{e}$, with $\bar{e} < 1000$, such that $U_0(\bar{e}) = U_{1000}(\bar{e})$. Furthermore, $U_0(e_i) > U_{1000}(e_i)$ for $e_i < \bar{e}$ and $U_0(e_i) < U_{1000}(e_i)$ for $1000 > e_i > \bar{e}$, that is, there is a threshold strategy.

First, note that $U_0(1000) = s(1000 - 1/\gamma) < U_{1000}(1000) = s(1000 - 1/\gamma) + 1$. Thus by continuity, types e_i close enough to 1000 will strictly prefer 1000 to the solution in (14). Then notice that $U_0(e_i)$ increases linearly in e_i with derivative s, while the derivative of U_{1000} with respect to e_i is

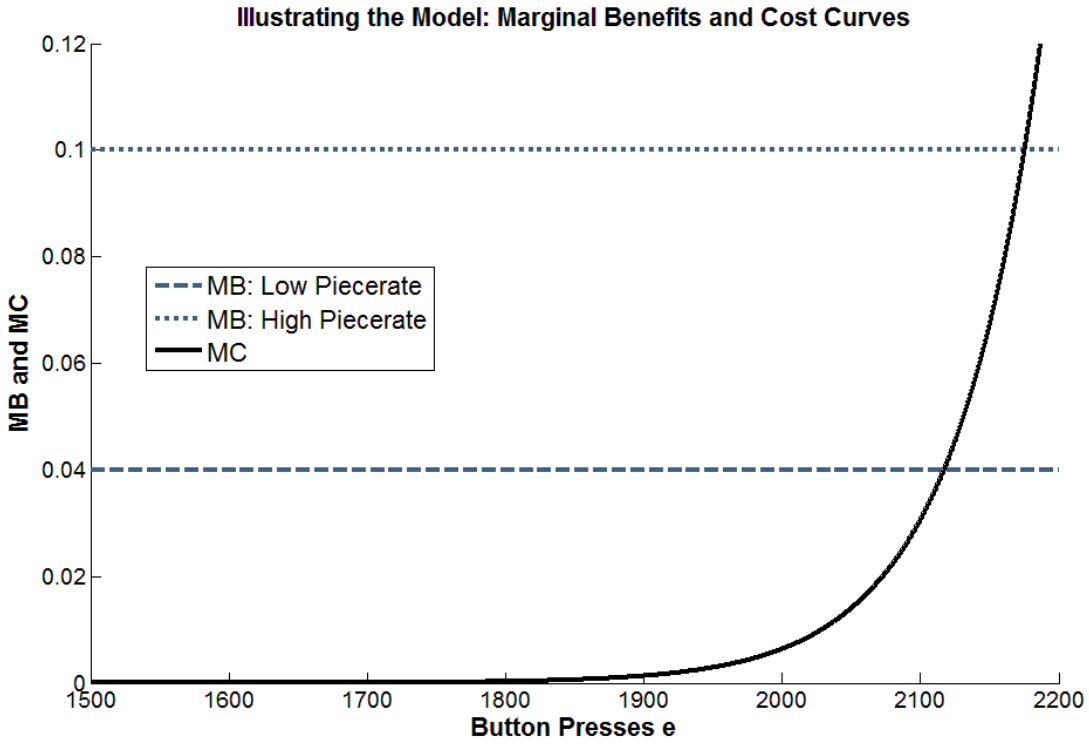
$$\frac{\partial U_{1000}(e_i)}{\partial e_i} = s\exp(\gamma(1000 - e_i)) > 0.$$

This derivative is decreasing in e_i, that is, U_{1000} is an increasing and concave function of e_i. Furthermore, for negative enough e_i the derivative $\partial U_{1000}(e_i)/\partial e_i$ becomes arbitrarily large and thus larger than the derivative $\partial U_0(e_i)/\partial e_i$. Thus, for e_i small enough, it must be the case that $U_0(e_i) > U_{1000}(e_i)$. To show that there is only one point of crossing between U_0 and U_{1000} for $e < 1000$, consider once again the properties of the two derivative functions. This concludes the proof.

Having determined the threshold $\bar{e}^{1000}$ we can similarly derive the other thresholds $\bar{e}^{2000}$ and $\bar{e}^{3000}$. Thus we know that the observed distribution will consist of a mixture of density from 0 to $\bar{e}^{1000}$, bunching at 1,000, then density from 1,000 to $\bar{e}^{2000}$ and so on. For the estimated $\hat{s}, \hat{\gamma}$, and $\hat{k}$, the threshold expressed in effort units are 185 (so types with effort higher than 185 and lower than 1,000 will jump to 1,000), 1,130, and 2,097.

The only remaining piece to determine is the distribution of the error term ε_i. We present the results following two approaches. The first approach just takes the estimated standard deviation of ε from the non-linear least squares estimation. The second approach instead backs out the distribution of ε non-parametrically from the no-payment case: an observed e_i, given the estimated $\hat{s}, \hat{\gamma}$, and $\hat{k}$, implies a realization of ε_i. Under either approach, we compute the counterfactual effort for the low-pay treatment, by moving the observations which are predicted to bunch to the bunching point, and then compute the expected effort. The first approach yields a simulated mean effort of 1,881, while the second approach yields a similar counterfactual of 1,878 for the mean effort. Thus, the effort is similar to the counterfactual estimated assuming continuous point earning.

Figure 1. Model of Effort Determination, Marginal Benefit and Marginal Cost



Notes: Figure 1 plots the determination of the equilibrium effort at the intersection of marginal cost and marginal benefit. The different piece rate treatments shift the marginal benefit curve, holding the marginal cost curve constant.

Figure 2a-b. Estimate of Model on 3 Benchmark Treatments

Figure 2a. Estimate with 0c, 1c, 10c Piece Rate and Prediction for 4c Piece Rate

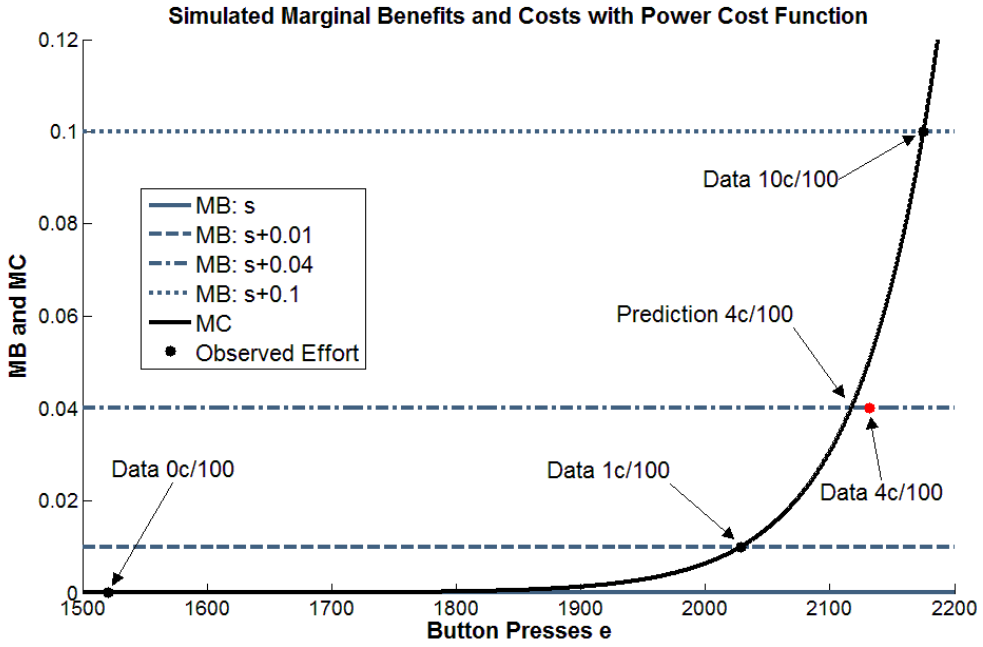
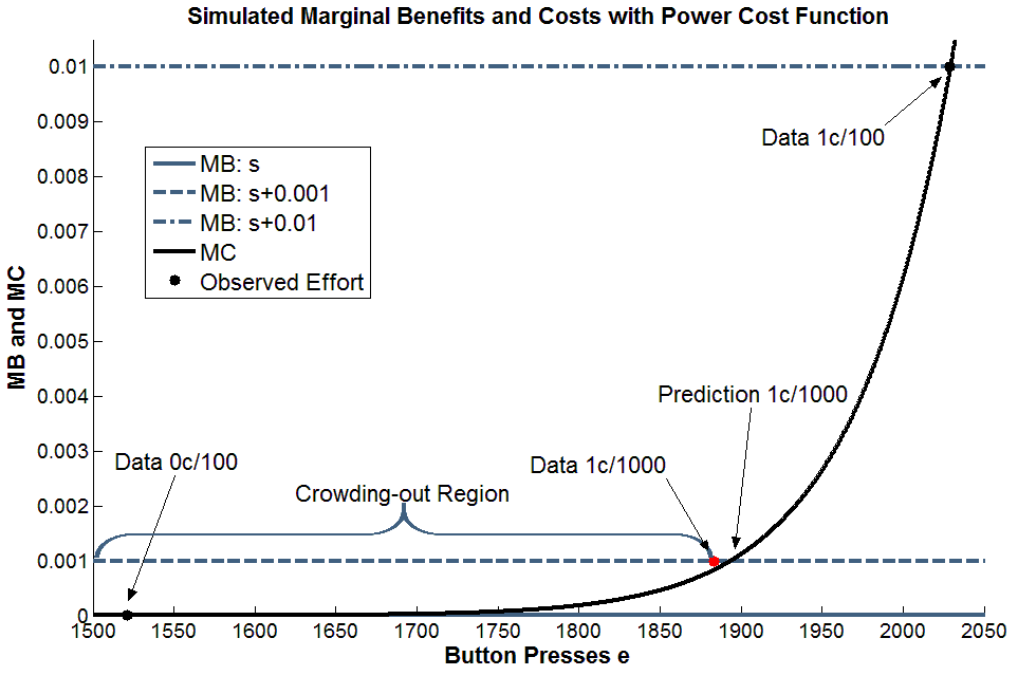
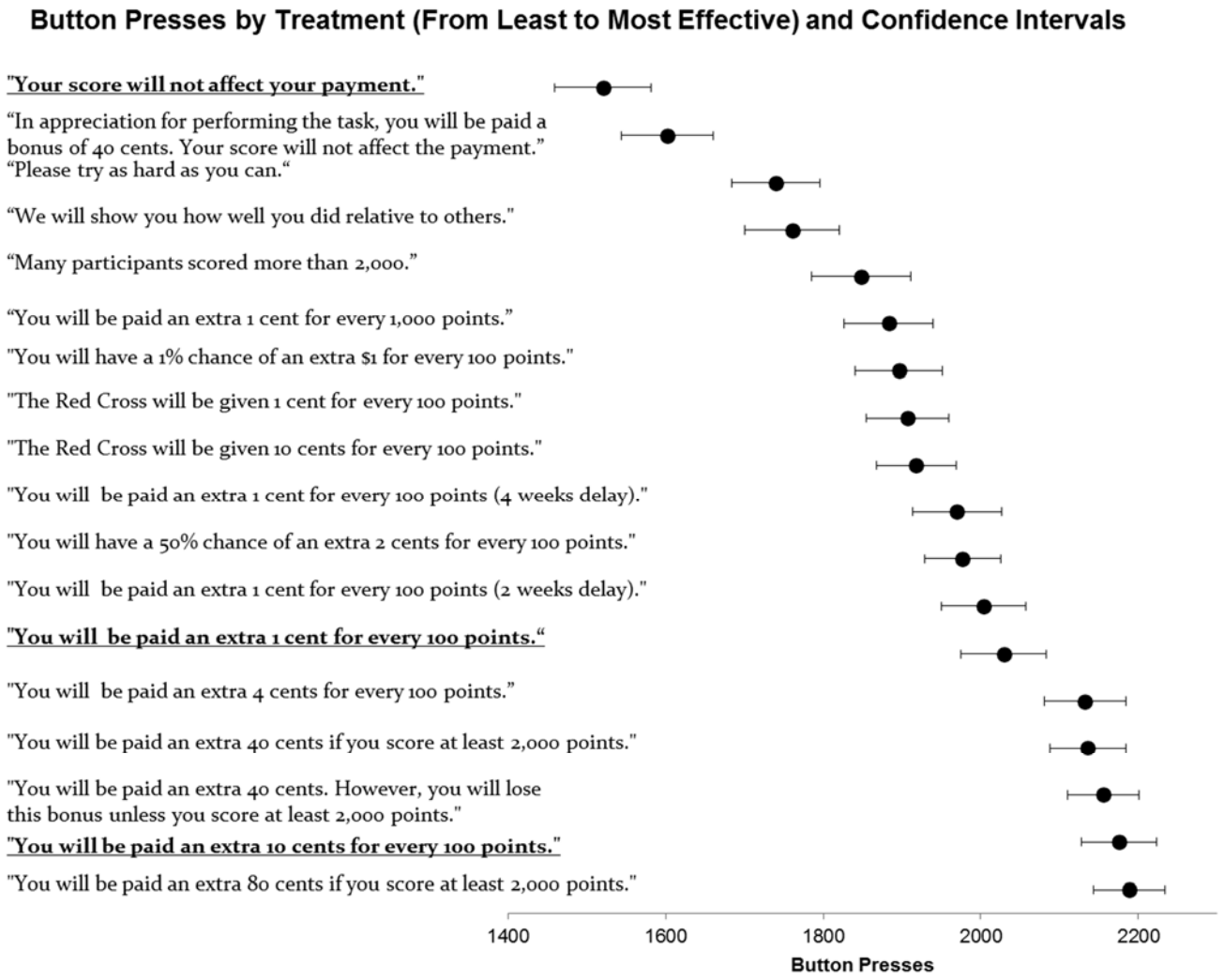


Figure 2b. Predicted Effort for "Paying Too Little" treatment (1 cent for 1,000 presses)



Notes: Figure 2a plots the marginal cost curve and the marginal benefit curve for the three benchmark treatments for the power cost function estimates. The marginal benefit curve equals the estimated s (warm glow) plus the piece rate. The marginal cost curve equals ke^s at the estimated k and s. At the estimates, we fit the three benchmark levels of effort perfectly, given that the model is just identified. Figure 2a also plots the out of sample prediction for the 4 cent treatment (which is not used in the estimates), as well as the observed effort for that treatment. Figure 2b plots, for the same point estimates, the out of sample prediction for the treatment with 1-cent per 1,000 clicks.

Figure 3. Average Button Presses by Treatment in Amazon Turk Task



Notes: Figure 3 presents the average score and confidence interval for each of 18 treatments in a real-effort task on Amazon Turk. Participants in the task earn a point by for each alternating a-b button press within a 10-minute period. The 18 treatments differ only in one paragraph presenting the treatments, the key sentence of which is reproduced in the first row. Each treatment has about 550 participants.

Figures 4a-c. Distribution of Effort, MTurk Workers, Cumulative Distribution Function

Figure 4a. Piece-Rate Treatments

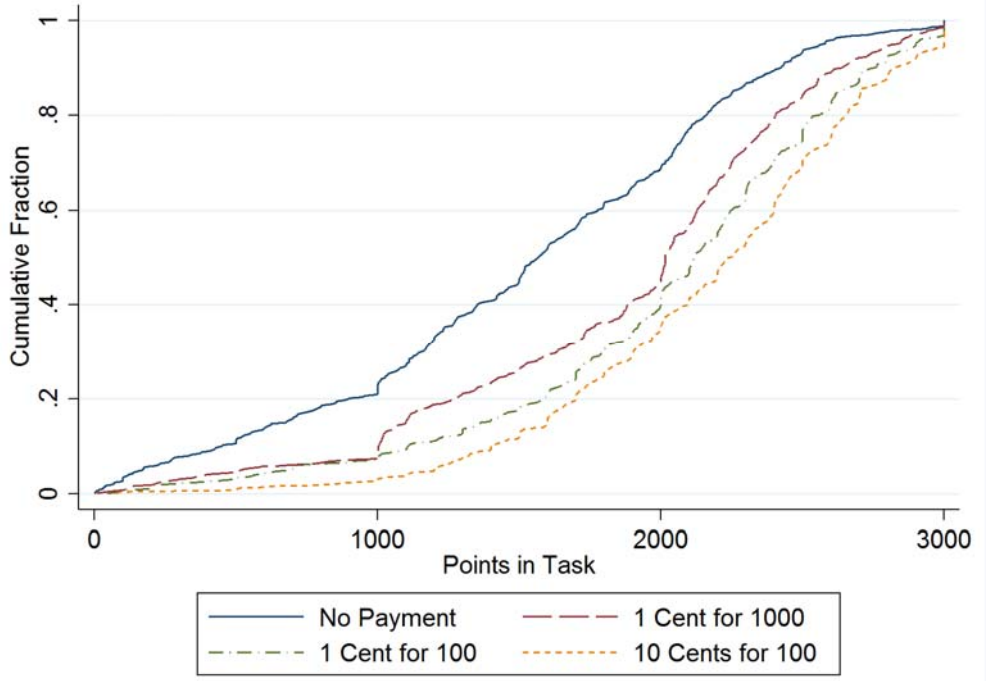


Figure 4b. Treatments with no monetary payoff

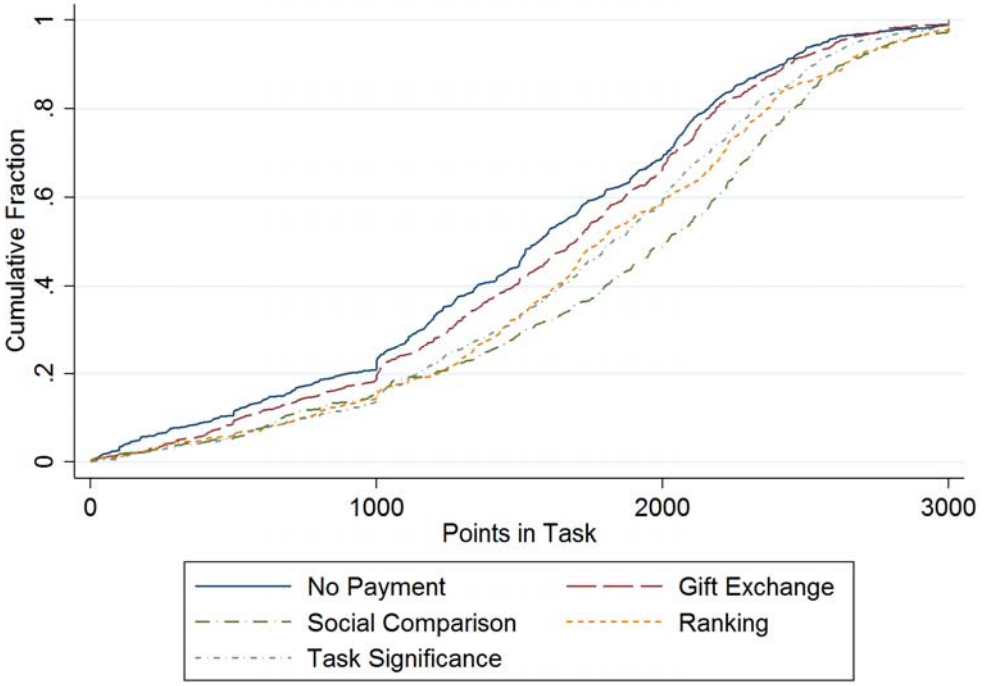
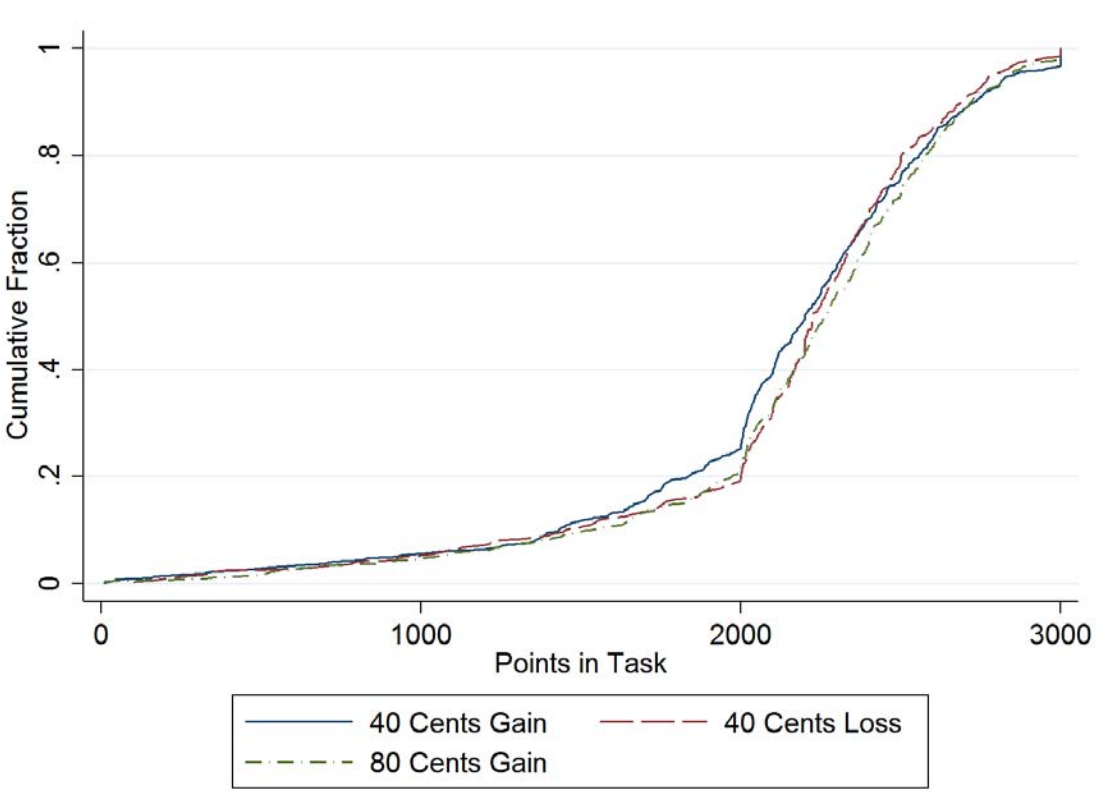


Figure 4c. Gain-Loss Treatments



Notes: Figures 4a-c present the cumulative distribution function of points for the MTurk workers in each of the treatments featured. The sample size in each treatment is approximately 550 subjects. Figure 4a features the three benchmark treatments (no piece rate, 1-cent per 100 points and 10 cents per 100 points), as well as the low-piece-rate treatment, 1 cent per 1,000 points. Figure 4b presents the results for the four treatments with no incentives (except for the charity treatments). Figure 4c presents the results for the gain-loss treatments.

Figure 5. Effort over Time, MTurk Workers

Figure 5a. Treatments with no Incentives and Piece Rate Treatments

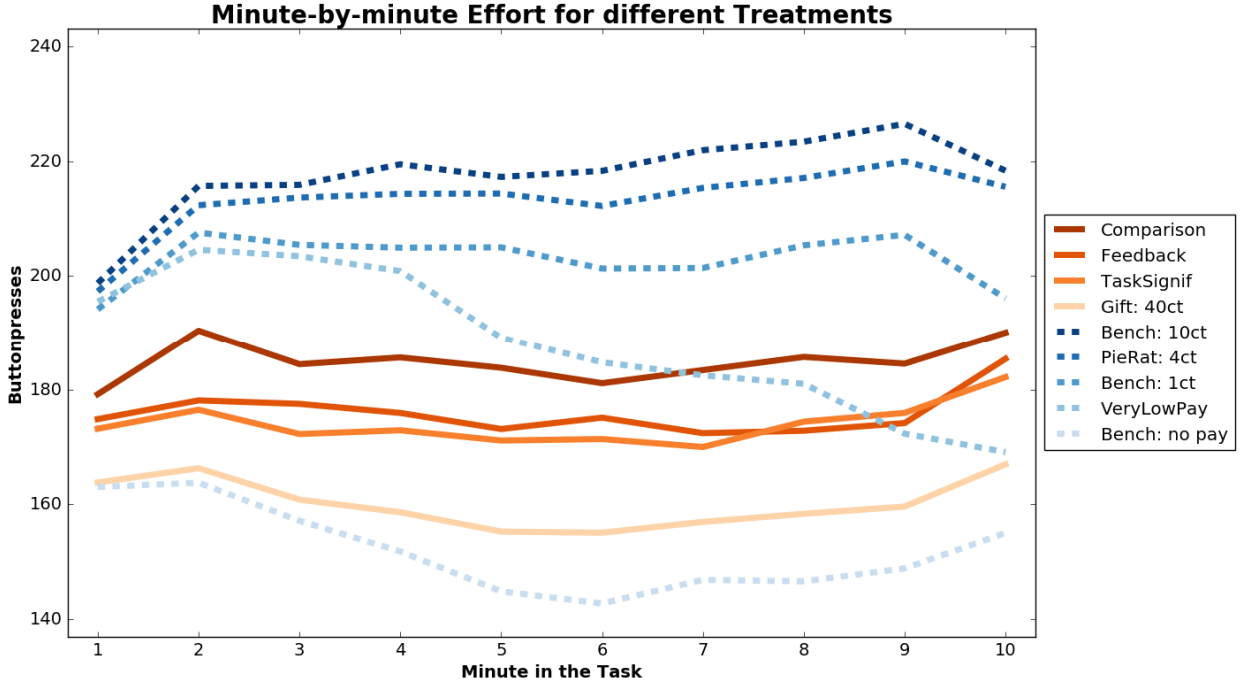
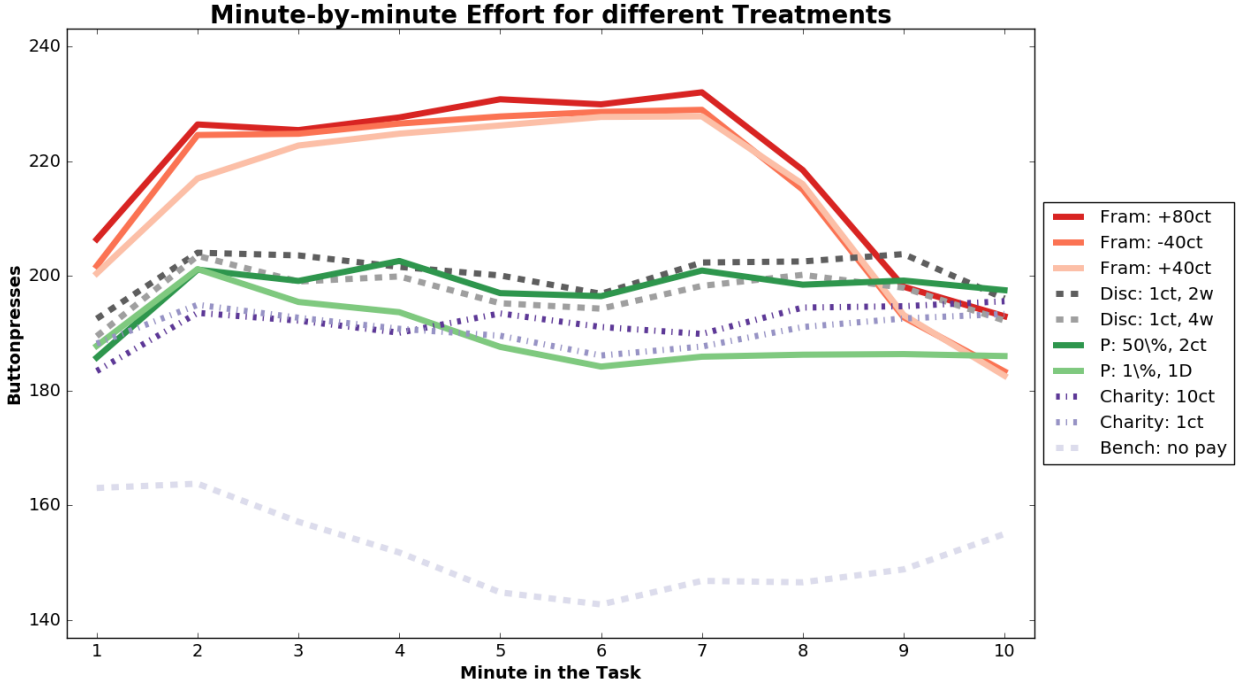


Figure 5b. Other Treatments



Notes: Figure 5 presents the effort over time for selected treatments. The y axis indicates the average number of button presses in that treatment per minute.

Figure 6. Average Button Presses by Treatment and Average Expert Forecasts

Notes: The black circles in Figure 6 present the average score for each of 18 treatments in a real-effort task on Amazon Turk. Participants in the task earn a point for each alternating a-b button press within a 10-minute period. The 18 treatments differ only in one paragraph presenting the treatments, the key sentence of which is reproduced in the first row. Each treatment has about 550 participants. The orange squares represent the average forecast from the sample of 208 experts who provided forecasts for the treatments. The three bolded treatments are benchmarks; the average score in the three benchmarks was revealed to the experts and thus there is no forecast.

Figures 7a-d. Heterogeneity of Expert Forecasts, Cumulative Distribution Function

Figure 7a. Piece-Rate and Charity Treatments

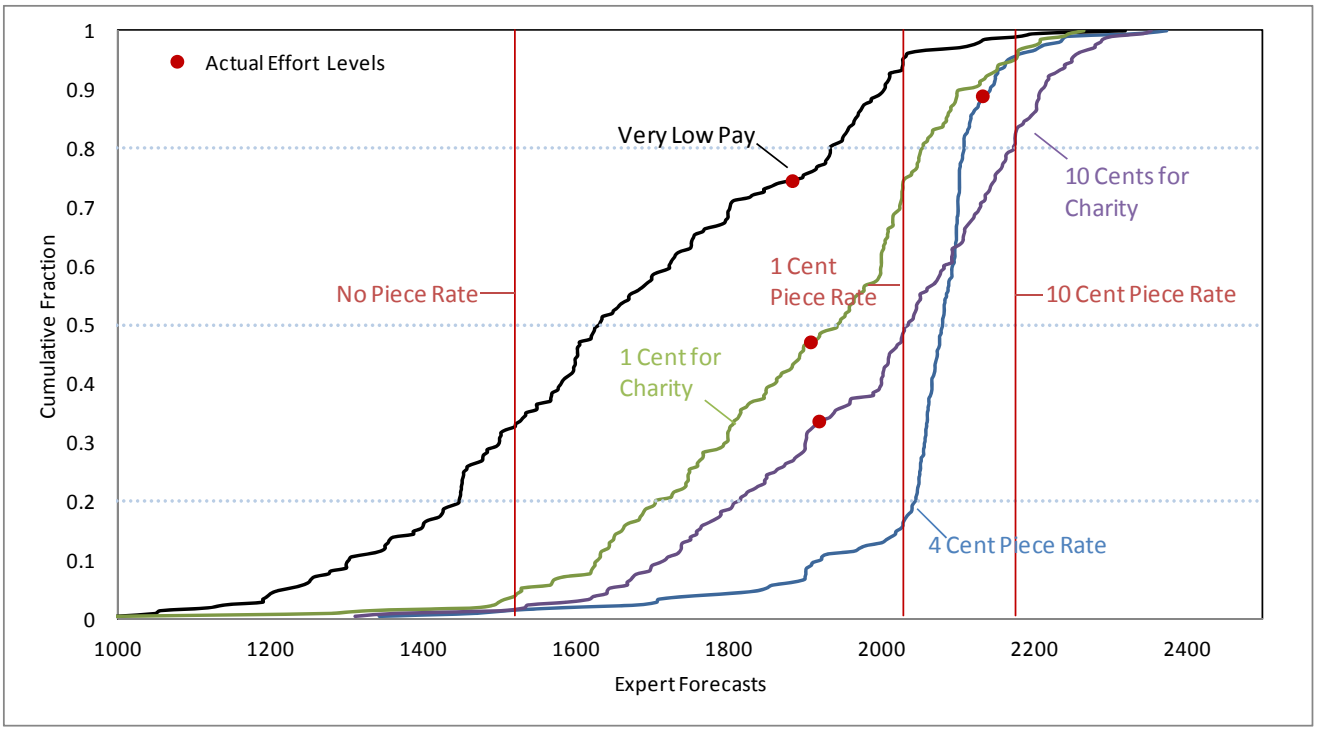


Figure 7b. Time Preference and Probability Weighting Treatments

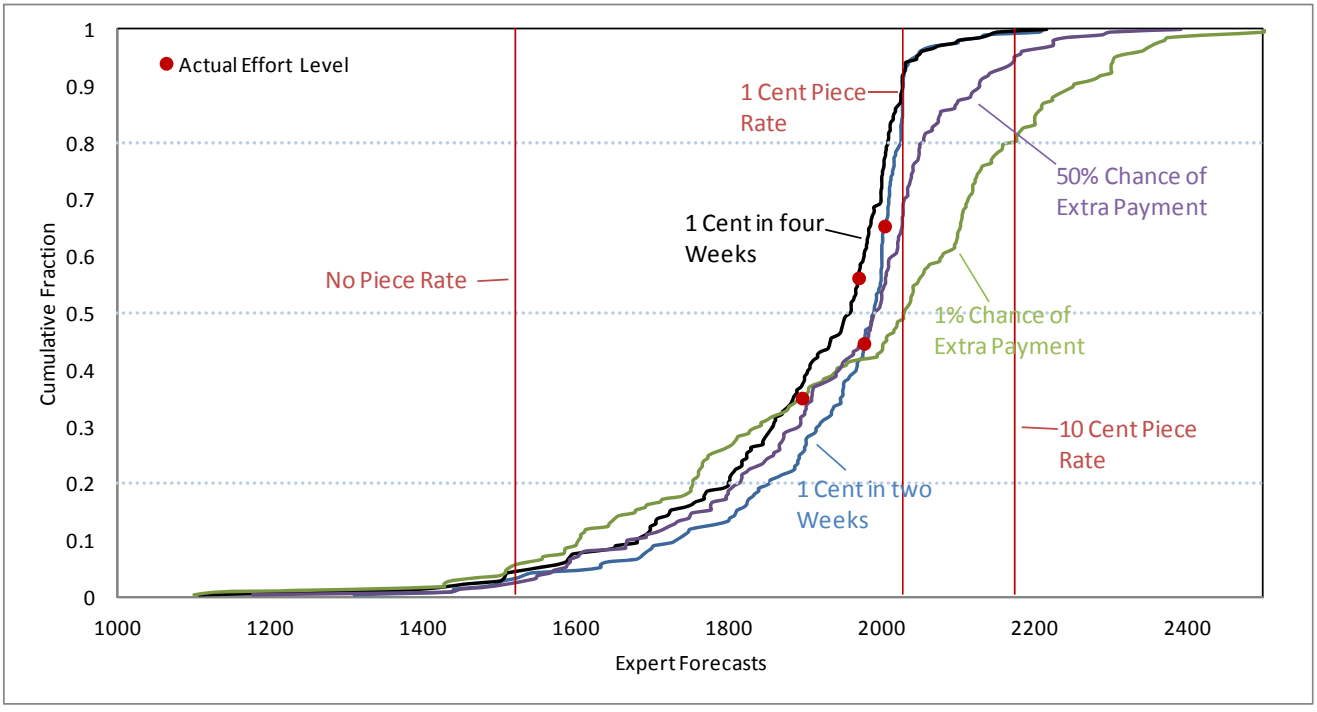


Figure 7c. Gain and Loss Treatments

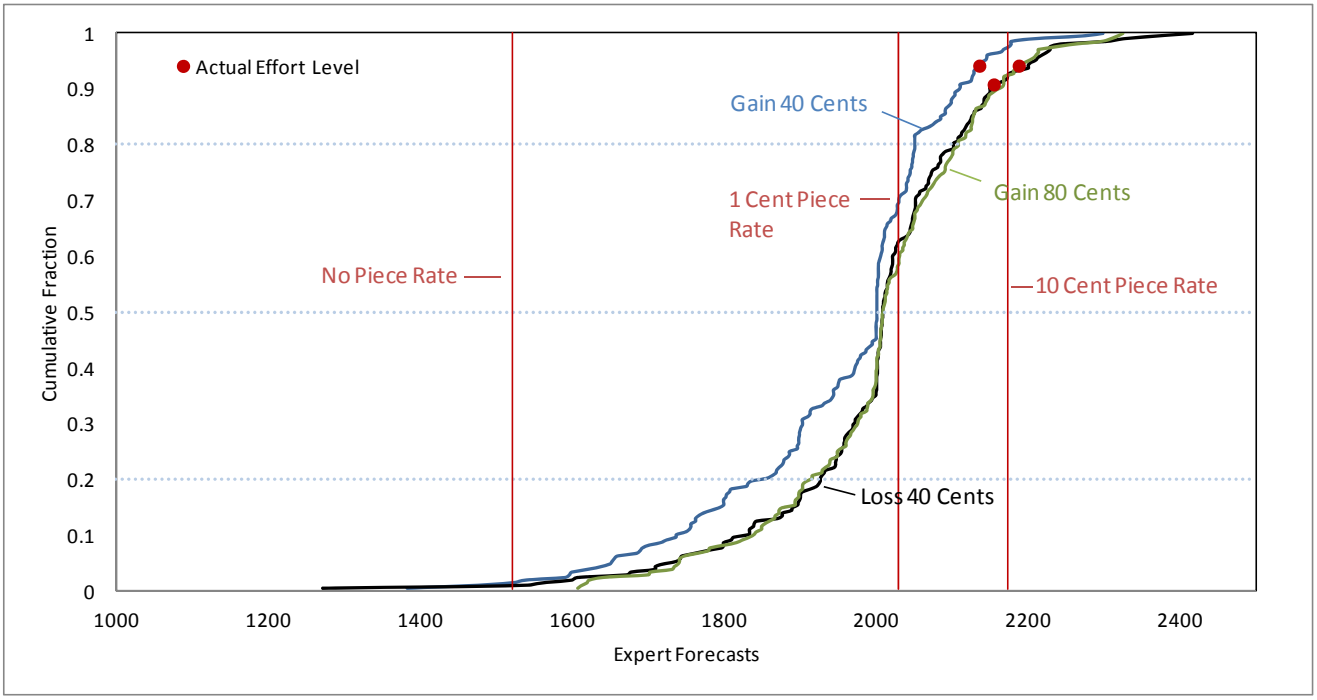
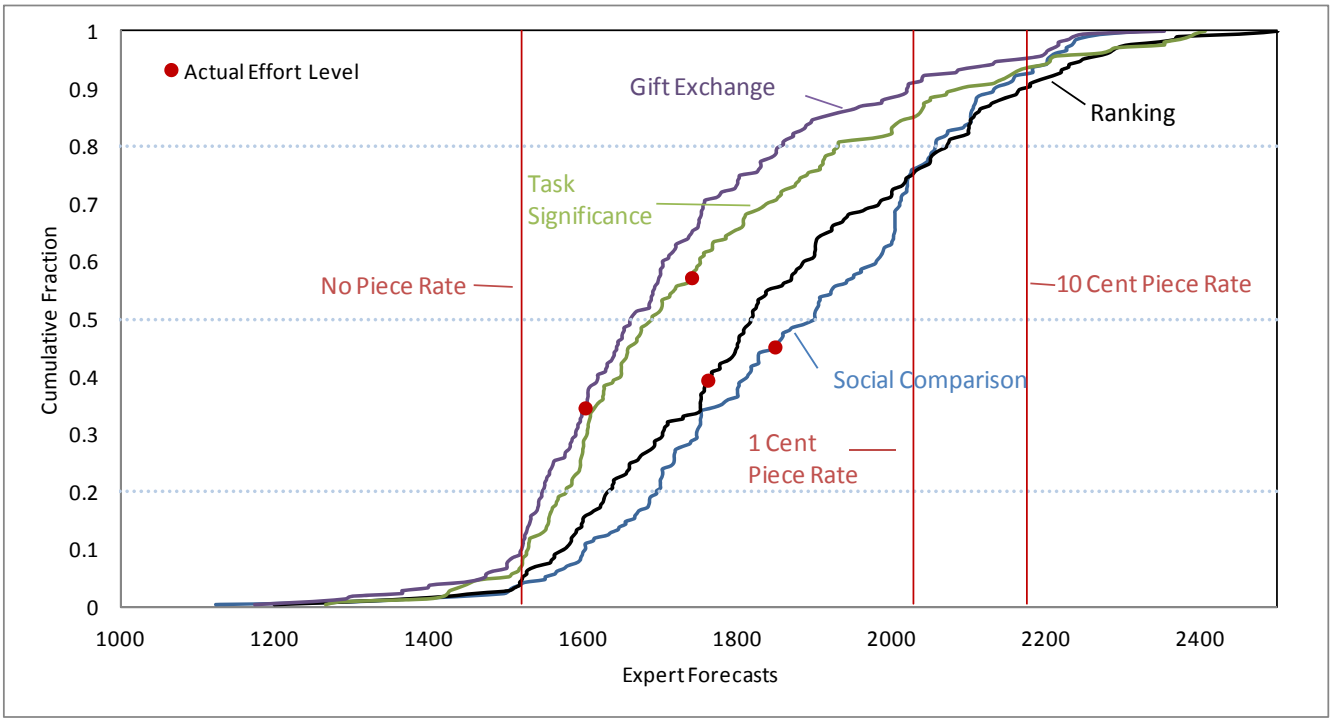
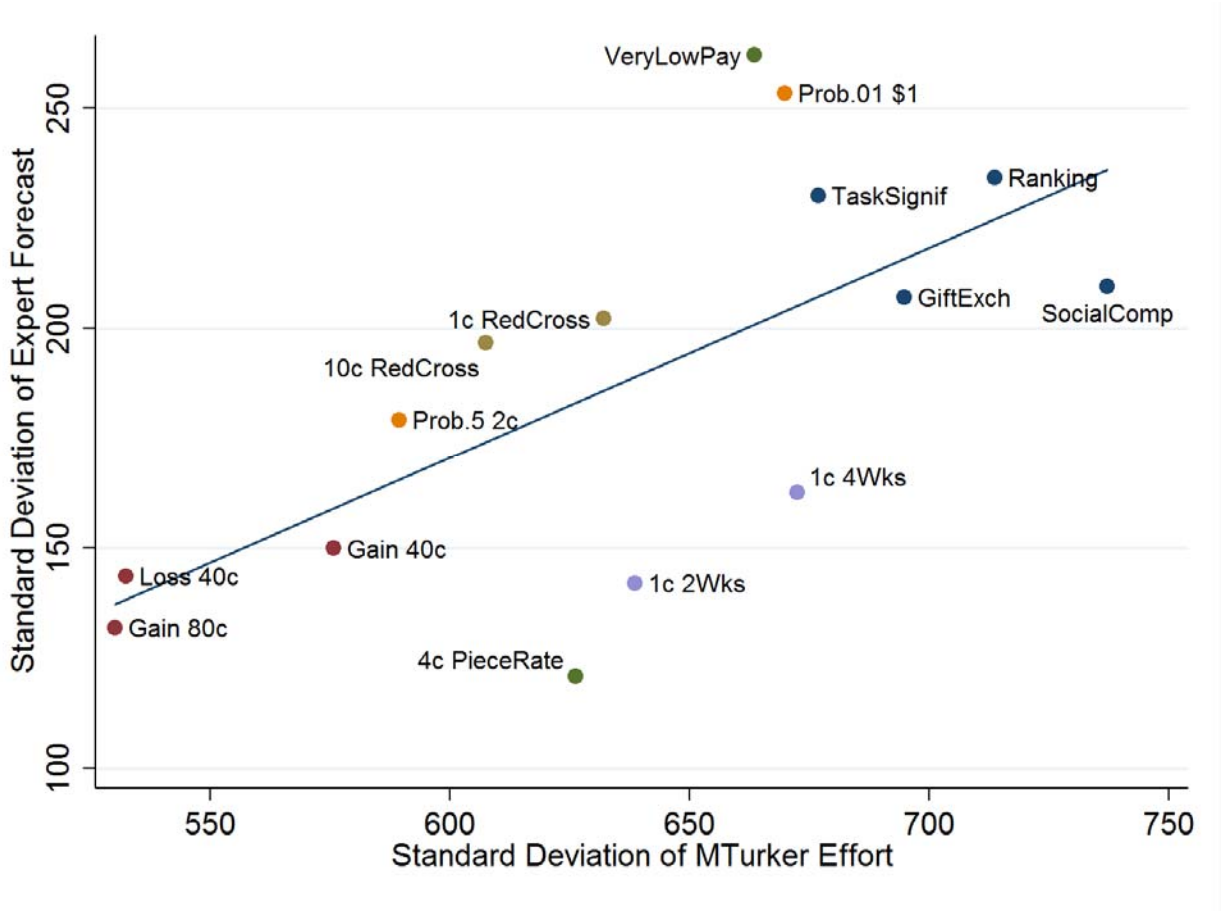


Figure 7d. Gift Exchange and Psychology Treatments



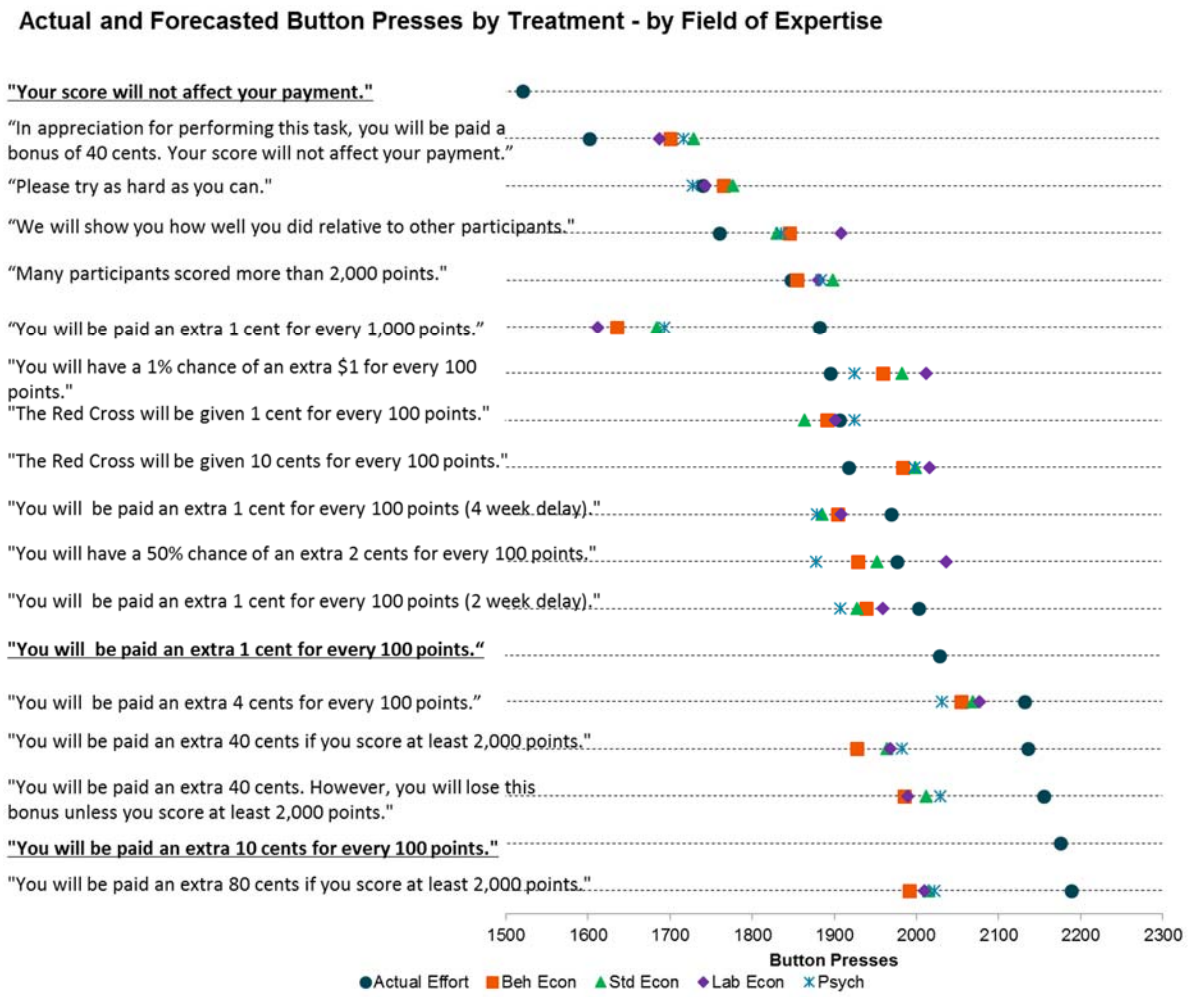
Notes: Figures 7a-d present the cumulative distribution function of forecasts by the 208 experts (see Table 1 for the list of treatment). The red circle presents the actual average score for that treatment. The vertical red lines present the score in the three benchmark treatments. Since the average score in the three benchmarks was revealed to the experts, there is no forecast for those.

Figures 8. Heterogeneity of Expert Forecasts and Heterogeneity of MTurker Effort, by Treatment



Notes: Figure 8 presents a scatterplot of the 15 treatments, with the standard deviation in MTurker effort on the x axis and the standard deviation in the expert forecast on the y axis. The figure also displays the best-fit line.

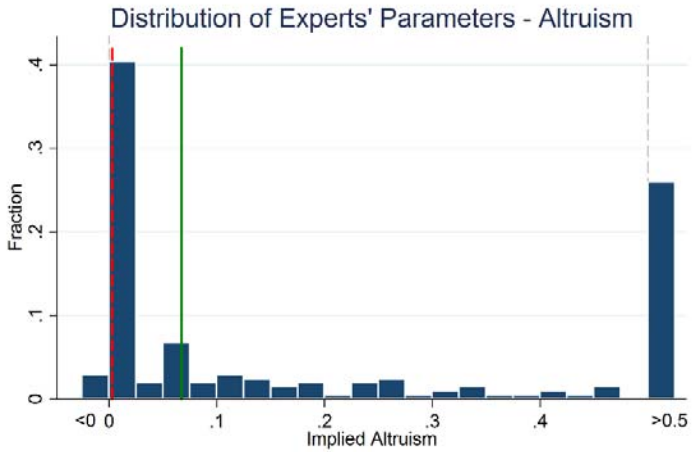
Figure 9. Average Button Presses by Treatment and Average Expert Forecasts, By Academic Field of Expert



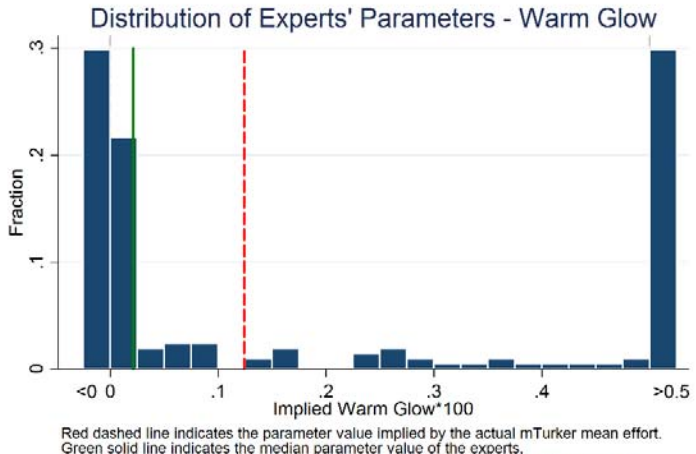
Notes: Figure 9 follows the same format of Figure 6, except that it splits the forecasts by the primary field of the 208 academic experts: behavioral economics, standard economics (consisting of applied microeconomics and economic theory) laboratory experiments, and psychology (which includes experts in behavioral decision-making).

Figure 10. Structural Estimates of Behavioral Parameters: Data versus Experts Beliefs

Figures 10a-b. Estimate of Social Preference Parameters

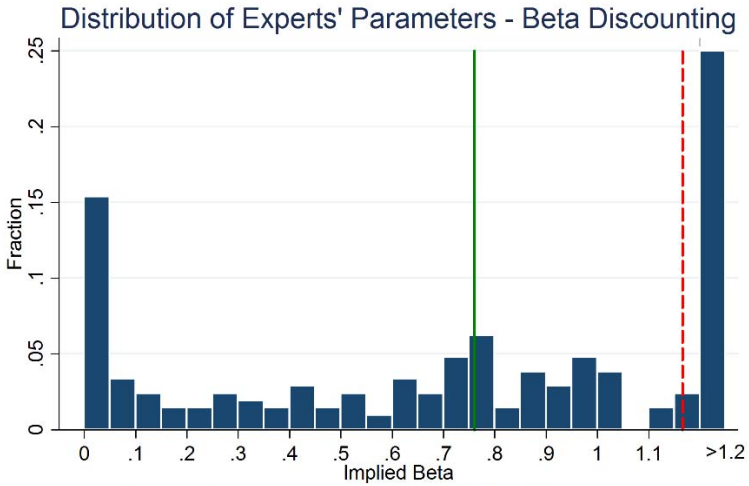


Red dashed line indicates the parameter value implied by the actual mTurker mean effort.
Green solid line indicates the median parameter value of the experts.
Observations are truncated at 0 and 0.5.

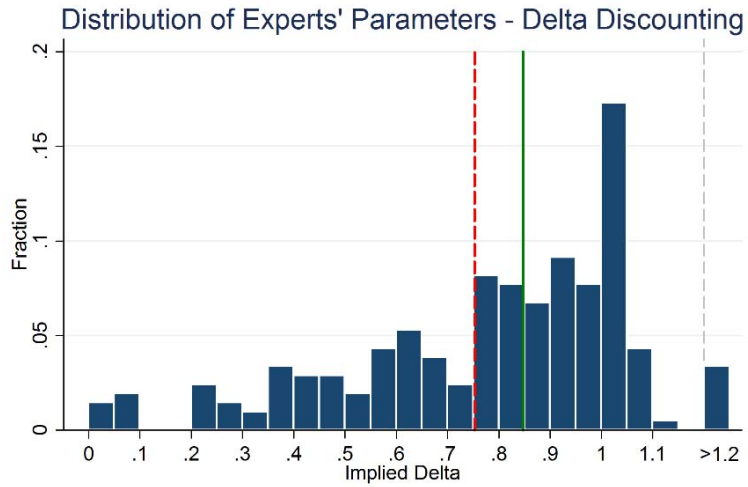


Red dashed line indicates the parameter value implied by the actual mTurker mean effort.
Green solid line indicates the median parameter value of the experts.
Parameter values are multiplied by 100 to faciliate interpretation as a an intrinsic value corresponding to paying an additional piece rate of x per 100 effort units.
Observations are truncated at 0 and 0.5.

Figures 10c-d. Estimate of Time Preference Parameters

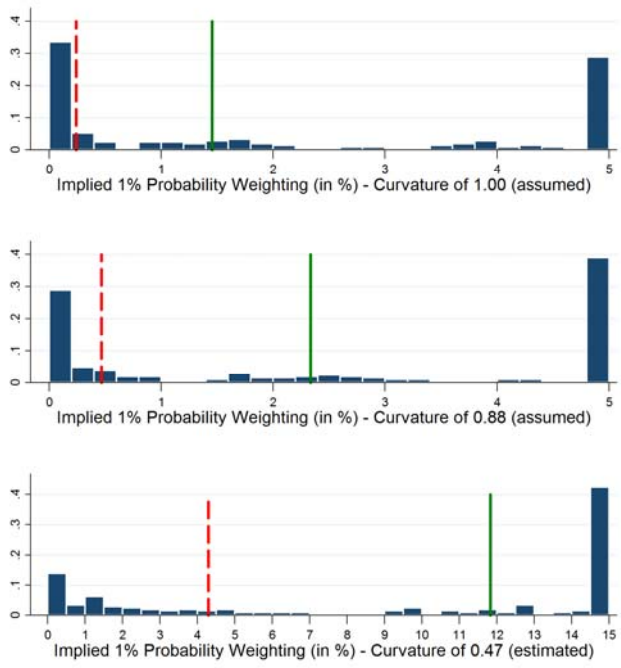


Red dashed line indicates the parameter value implied by the actual mTurker mean effort.
Green solid line indicates the median parameter value of the experts.
Observations are truncated at 1.2.

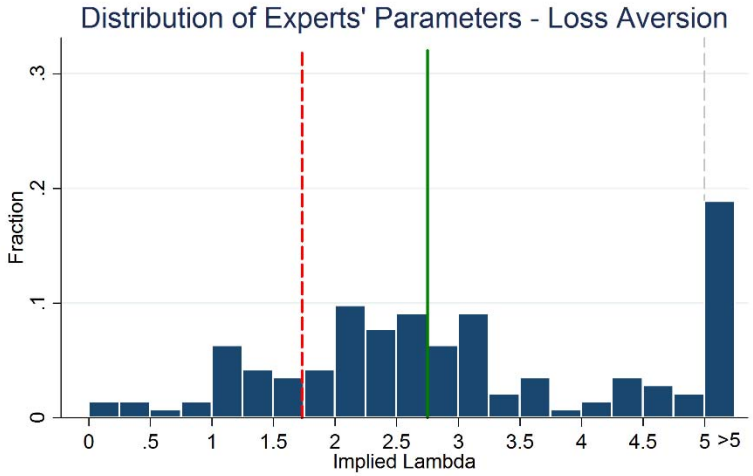


Red dashed line indicates the parameter value implied by the actual mTurker mean effort.
Green solid line indicates the median parameter value of the experts.
Observations are truncated at 1.2

Figures 10e-f. Estimate of Reference-Dependence Parameters



Implied probability weighting under three different specifications of the curvature over the piece rate (1, 0.88, 0.47). The red line indicates the actual probability weighting from the Mturkers' effort, the green line denotes the median forecast of the experts. Observations are truncated at 5, 5, and 15, respectively.



Red dashed line indicates the parameter value implied by the actual mTurker mean effort.
Green solid line indicates the median parameter value of the experts.
Observations are truncated at 5. Observations with a difference of less than 10 in the two gain-treatments and negative parameter values are dropped (i.e., 58 and 7 observations)

Notes: Figures 10a-f present the distribution of the estimates of the behavioral parameters from the relevant treatments (see Table 5). We use a minimum-distance estimator to estimate a model of costly effort with a power cost of effort function using the average effort in the three benchmark treatments for Figures 10a-d. The resulting parameter estimates are in Column (1), Panel A of Table 5. For Figure 10e we use a non-linear least squares estimate with an exponential cost function as in Table 6, Columns 4-6. Figure 10f is based on an approximate solution (see text). We use these estimated parameters and the observed effort in the relevant treatments to back out the implied structural estimate for a behavioral parameter from the relevant treatment (plotted as the red vertical line). Similarly, for each expert *i* we back out the expected behavioral parameter implied by the forecast which expert *i* makes for a particular treatment; the implied structural parameters are plotted in the figures, with the green line denotes the median parameter. See also the results in Panel B of Table 5. Figures 10a-b plot the implied altruism and warm glow parameters from the charitable giving treatments. Figure 10c-d plot the implied *beta* and *delta* from the time preference treatments. Figures 10e-f plot the implied probability weight (corresponding to a .01 probability) and loss aversion from the reference dependence treatments.

Table 1: Summary of 18 Treatments

Category	Treatment Wording	Parameter	Cites
(1)	(2)	(3)	(4)
Piece Rate	"Your score will not affect your payment in any way."		
	As a bonus, you will be paid an extra **1 cent** for every **100 points** that you score."		
	"As a bonus, you will be paid an extra **10 cents** for every **100 points** that you score."		
	"As a bonus, you will be paid an extra **4 cents** for every **100 points** that you score."		
Pay Enough or Don't Pay	"As a bonus, you will be paid an extra **1 cent** for every **1,000 points** that you score."	ΔS_{CO} (crowd out)	Deci, 1971; Gneezy and Rustichini, 2000
Social Preferences: Charity	"As a bonus, the Red Cross charitable fund will be given **1 cent** for every 100 points that you score."	α (altruism) and *a* (warm glow)	Andreoni, 1989 and 1990; Becker, 1972; Imas, 2014
	"As a bonus, the Red Cross charitable fund will be given **10 cents** for every 100 points that you score."		
Social Preferences: Gift Exchange	"In appreciation to you for performing this task, you will be paid a **bonus** of **40 cents**. Your score will not affect your payment in any way."	ΔS_{GE}	Fehr, Kirchsteiger, and Riedl, 1993; Gneezy and List, 2009
Discounting	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **two weeks** from today."	β, δ (impatience parameters)	Laibson, 1997; O'Donoghue and Rabin, 1999; Andreoni and Sprenger, 2012; Augenblick, Nierderle, and Sprenger, 2015
	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **four weeks** from today."		
Gains versus Losses	"As a bonus, you will be paid an **extra 40 cents** if you score at least 2,000 points."	λ (loss aversion)	Kahneman and Tversky, 1979; Hossain and List, 2012; Fryer, Levitt, List, Sadoff, 2012
	"As a bonus, you will be paid an **extra 40 cents**. However, you will **lose this bonus** (it will not be placed in your account) unless you score at least 2,000 points. "		
	"As a bonus, you will be paid an **extra 80 cents** if you score at least 2,000 points."		
Risk Aversion and Probability Weighting	"As a bonus, you will have a **1% chance** of being paid an **extra $1** for every 100 points that you score. One out of every 100 participants who perform this task will be randomly chosen to be paid this reward."	π(P) (probability weighting)	Kahneman and Tversky, 1979; Prelec, 1998; Wu and Gonzalez, 1996; Loewenstein, Brennan, and Volpp, 2007
	"As a bonus, you will have a **50% chance** of being paid an **extra 2 cents** for every 100 points that you score. One out of two participants who perform this task will be randomly chosen to be paid this reward."		
Social Comparisons	"Your score will not affect your payment in any way. In a previous version of this task, **many participants** were able to score more than **2,000 points.**"	ΔS_{SC}	Goldstein, Cialdini, and Griskevicius, 2008
Ranking	"Your score will not affect your payment in any way. After you play, we will show you **how well you did relative** to other participants who have previously done this task."	ΔS_{R}	Maslow, 1943; Bandiera et al., 2013; Ashraf et al., 2012
Task Significance	"Your score will not affect your payment in any way. We are interested in how fast people choose to press digits and we would like you to do your very best. So **please try as hard as you can.**"	ΔS_{TS}	Grant, 2008

Notes: The Table lists the 18 treatments in the Mturk experiment. The treatments differ just in one paragraph explaining the task and in the vizualization of the points earned. Column (2) reports the key part of the wording of the paragraph. For brevity, we omit from the description the sentence "This bonus will be paid to your account within 24 hours" which applies to all treatments with incentives other than in the Time Preference ones where the payment is delayed. Notice that the bolding is for the benefit of the reader of the Table and was not used in the treatment description on MTurk. Column (1) reports the conceptual grouping of the treaments, Column (3) reports the parameters in the model related to the treatment, and Column (4) reports some key references for the treatment.

Table 2. Summary Statistics, Mturk Sample

	Mean	US Census
	(1)	(2)
Button Presses	1936	
Time to complete survey (minutes)	12.90	
US IP Address Location	0.85	
India IP Address Location	0.12	
Female	0.54	0.52
Education		
High School or Less	0.09	0.44
Some College	0.36	0.28
Bachelor's Degree or more	0.55	0.28
Age		
18-24 years old	0.21	0.13
25-30 years old	0.30	0.10
31-40 years old	0.27	0.17
41-50 years old	0.12	0.18
51-64 years old	0.08	0.25
Older than 65	0.01	0.17
Observations	9861	

Notes: Column (1) of Table 2 lists summary statistics for the final sample of Amazon Turk survey participants (after screening out ineligible subjects). Column (2) lists, where available, comparable demographic information from the US Census.

Table 3. Summary Statistics, Experts

	All Experts Contacted	Experts Completed Survey	Experts Completed All 15 Treatments
	(1)	(2)	(3)
Primary Field			
Behavioral Econ.	0.25	0.31	0.32
Behavioral Finance	0.06	0.05	0.04
Applied Micro	0.17	0.19	0.19
Economic Theory	0.09	0.07	0.07
Econ. Lab Exper.	0.17	0.15	0.16
Decision Making	0.17	0.12	0.13
Social Psychology	0.08	0.10	0.10
Academic Rank			
Assistant Professor	0.26	0.36	0.36
Associate Professor	0.15	0.15	0.15
Professor	0.55	0.45	0.45
Other	0.04	0.04	0.04
Minutes Spent (med.)			17
Clicked Practice Task			0.44
Clicked Instructions			0.22
Heard of Mturk			0.98
Used Mturk			0.51
Observations	314	213	208

Notes: The Table presents summary information on the experts participating in the survey. Column (1) presents information on the experts contacted and Column (2) on the experts that completed the survey. Column (3) restricts the sample further to subjects who made a forecast for all 15 treatments.

Table 4. Findings by Treatment: Effort in Experiment and Expert Forecasts

Category	Treatment Wording	N	Mean Effort (s.e.)	Mean Forecast	Std. Dev. Forecast	Actual - Forecast (s.e.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Piece Rate	"Your score will not affect your payment in any way."	540	1521 (31.22)		Benchmark	
	As a bonus, you will be paid an extra **1 cent** for every **100 points** that you score."	558	2029 (27.47)		Benchmark	
	"As a bonus, you will be paid an extra **10 cents** for every **100 points** that you score."	566	2175 (24.29)		Benchmark	
	"As a bonus, you will be paid an extra **4 cents** for every **100 points** that you score."	562	2132 (26.41)	2057	120.86	75 (27.71)
Pay Enough or Don't Pay	"As a bonus, you will be paid an extra **1 cent** for every **1,000 points** that you score."	538	1883 (28.61)	1657	262.00	226 (33.89)
Social Preferences: Charity	"As a bonus, the Red Cross charitable fund will be given **1 cent** for every 100 points that you score."	554	1907 (26.86)	1894	202.20	13 (30.30)
	"As a bonus, the Red Cross charitable fund will be given **10 cents** for every 100 points that you score."	549	1918 (25.93)	1997	196.75	-79 (29.30)
Social Preferences: Gift Exchange	"In appreciation to you for performing this task, you will be paid a **bonus** of **40 cents**. Your score will not affect your payment in any way."	545	1602 (29.77)	1709	207.12	-107 (33.05)
Discounting	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **two weeks** from today."	544	2004 (27.38)	1933	142.02	71 (29.10)
	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **four weeks** from today."	550	1970 (28.68)	1895	162.54	75 (30.81)
Gains versus Losses	"As a bonus, you will be paid an **extra 40 cents** if you score at least 2,000 points."	545	2136 (24.66)	1955	149.90	181 (26.76)
	"As a bonus, you will be paid an **extra 40 cents**. However, you will **lose this bonus** (it will not be placed in your account) unless you score at least 2,000 points. "	532	2155 (23.09)	2002	143.57	153 (25.14)
	"As a bonus, you will be paid an **extra 80 cents** if you score at least 2,000 points."	532	2188 (22.99)	2007	131.93	181 (24.74)
Risk Aversion and Probability Weighting	"As a bonus, you will have a **1% chance** of being paid an **extra $1** for every 100 points that you score. One out of every 100 participants who perform this task will be randomly chosen to be paid this reward."	555	1896 (28.44)	1967	253.43	-71 (33.43)
	"As a bonus, you will have a **50% chance** of being paid an **extra 2 cents** for every 100 points that you score. One out of two participants who perform this task will be randomly chosen to be paid this reward."	568	1977 (24.73)	1941	179.27	36 (27.68)
Social Comparisons	"Your score will not affect your payment in any way. In a previous version of this task, **many participants** were able to score more than **2,000 points.**"	526	1848 (32.14)	1877	209.48	-29 (35.27)
Ranking	"Your score will not affect your payment in any way. After you play, we will show you **how well you did relative** to other participants who have previously done this task."	543	1761 (30.63)	1850	234.28	-89 (34.67)
Task Significance	"Your score will not affect your payment in any way. We are interested in how fast people choose to press digits and we would like you to do your very best. So **please try as hard as you can**."	554	1740 (28.76)	1757	230.15	-17 (32.89)

Notes: The Table lists the 18 treatments in the Mturk experiment. The treatments differ just in one paragraph explaining the task and in the vizualization of the points earned. Column (2) reports the key part of the wording of the paragraph. For brevity, we omit from the description the sentence "This bonus will be paid to your account within 24 hours" which applies to all treatments with incentives other than in the Time Preference ones where the payment is delayed. Notice that the bolding is for the benefit of the reader of the Table. In the actual description to the MTurk workers, the whole paragraph was bolded and underlined. Column (1) reports the conceptual grouping of the treaments, Columns (3) and (4) report the number of MTurk subjects in that treatment and the mean number of points, with the standard errors. Column (5) reports the mean forecast among the 208 experts of the points in that treatment. Column (6) reports the standard deviation among the expert forecasts for that treatment. Column (6) reports the difference between the average forecast and the actual average effort, with its standard errror.

Table 5. Estimates of Behavioral Parameters I: Mturkers Actual Effort and Expert Beliefs

Cost of Effort Specification:	Power Cost of Effort		Exponential Cost of Effort	
Estimation Method:	Minimum Distance Estimator on Average Effort	Non-Linear Least Squares on Individual Effort	Minimum Distance Estimator on Average Effort	Non-Linear Least Squares on Individual Effort
	(1)	(2)	(3)	(4)
Panel A. Estimate of Model on Effort in 3 Benchmark Treatments				
Curvature γ of Cost of Effort Function	33.21 (11.86)	24.07 (6.18)	0.0158 (0.0056)	0.0156 (0.0040)
Level *k* of Cost of Effort Function	1.46E-112 (2.25E-65)	6.54E-82 (.)	1.27E-16 (1.18E-11)	1.70E-16 (1.65E-13)
Intrinsic Motivation *s* (cent per 100 points)	6.96E-05 (1.06E-03)	8.35E-07 (2.46E-6)	3.32E-04 (2.45E-03)	3.69E-04 (7.97E-04)
Sum of Squared Errors	7.62E-05		2.92E-10	
R Squared		0.1331		0.1532
N	1664	1664	1664	1664
Implied Effort, 4-cent Treatment (Actual Effort 2,132, Log 7.602)	2116	7.586 (Expected log effort)	2117	2121
Implied Effort, Low-pay Treatment (Actual Effort 1,883, Log 7.424)	1893	7.413 (expected log effort)	1883	1885 / 1881 / 1878

Panel B. Estimates of Social Preferences and Time Preferences

	Estimate from Mturk (95% c.i.)	Median Forecast (25th, 75th ptile)	Estimate from Mturk (95% c.i.)	Estimate from Mturk (95% c.i.)	Median Forecast (25th, 75th ptile)	Estimate from Mturk (95% c.i.)	Median Forecast (25th, 75th ptile)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Social Preferences Parameters							
Pure Altruism Coefficient α	0.003 (-0.02, 0.03)	0.067 (0.002,0.548)	0.010 (-0.028,0.049)	0.003 (-0.02, 0.03)	0.067 (0.002,0.543)	0.004 (-0.018,0.025)	0.070 (0.002,0.539)
Warm Glow Coefficient *a* (cent per 100 points)	0.124 (0.00, 0.55)	0.020 (-0.001,0.736)	0.200 (-0.203,0.603)	0.143 (0.00, 0.56)	0.029 (0.000,0.705)	0.142 (-0.138,0.422)	0.034 (0.000,0.724)
Gift Exchange Δ*s* (cent per 100 points)	3.20E-04 (3.5E-9, 0.007)	0.001 (1.0E-4,0.022)	0.002 (-0.006, 0.009)	8.59E-04 (1.7E-8, 0.012)	0.003 (3.1E-4,0.031)	0.001 (-0.005, 0.008)	0.003 (3.3E-4,0.039)
Time Preference Parameters							
Present Bias β	1.17 (0.09, 9.03)	0.76 (0.27,1.22)	1.49 (-1.83,4.82)	1.15 (0.09, 8.40)	0.76 (0.29,1.19)	1.24 (-1.35,3.82)	0.79 (0.30,1.23)
(Weekly) Discount Factor δ	0.75 (0.34, 1.49)	0.85 (0.61,1.00)	0.73 (0.23,1.23)	0.76 (0.35, 1.45)	0.85 (0.64,1.00)	0.75 (0.26,1.27)	0.86 (0.64,1.00)

Notes: Panel A reports the structural estimates of the model in Section 2. Columns (1) and (3) use a minimum-distance estimator employing 3 moments (average effort in three benchmark treatments) and 3 parameters, and is thus exactly idenitified. We estimate the model under two assumptions, a power cost of effort function (Column (1)) and an exponential cost of effort function (Column (3)). The standard errors are derived via a bootstrap with 1,000 draws. Columns (2) and (4) use a non-linear least sqaures specification using the individual effort of MTurkers (rounded to the nearest 100) in the 3 benchmark treatments. Panel B uses the estimated model parameters to back out the implied estimates for the behavioral parameters. The confidence intervals for the minimum distance estimates are derived from the bootstrap. In the rows displaying the implied effort we compute the predicted effort given the parameters for the 4-cent treatment and the low-pay treatment. For the low-pay treatment in Column 4, we present two alternative predictions which explicitly model the discontinuity in payoffs, with very similar results (Appendix A). Columns (1), (3), (4), and (6) use the observed average effort in the relevant treatments to back out the parameters. Columns (2), (5), and (7) instead use the expert forecasts, showing the median, the 25th percentile and the 75th percentile of the parameters implied by the forecasts. We do not elicit parameters for the experts under the power cost function for the non-linear least squaress estimate since we did ask for the expected log effort, which is the key variable for that model.

Table 6. Estimates of Reference-Dependent Parameters: Mturker Actual Effort and Expert Beliefs

Estimation Method:	**Non-Linear Least Squares on Individual Effort in 3 Treatments**					
Cost of Effort Specification:	**Power Cost of Effort**			**Exponential Cost of Effort**		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Estimate of Model on Effort in 3 Benchmark Treatments and 2 Probability Treatments						
Curvature γ of Cost of Effort Function	20.59 (4.22)	18.87 (3.92)	19.64 (14.19)	0.0134 (0.0024)	0.0119 (0.0021)	0.0072 (0.0027)
Level *k* of Cost of Effort Function	3.38E-70 (5.45E-68)	3.92E-64 (1,16E-62)	1.02E-66 (1.12E-64)	2.42E-14 (1.19E-13)	7.50E-13 (3.27E-12)	5.46E-08 (3.50E-7)
Intrinsic Motivation *s* (cent per 100 points)	2.66E-04 (5.45E-4)	6.22E-04 (0.001)	3.75E-04 (0.003)	0.002 (0.002)	0.006 (0.007)	0.314 (0.716)
Probability Weighting π(.01) (in percent)	0.19 (0.15)	0.38 (0.26)	0.30 (1.31)	0.24 (0.14)	0.47 (0.24)	4.30 (5.25)
Curvature of Utility Over Piece Rate	1.00 (assumed)	0.88 (assumed)	0.92 (0.79)	1.00 (assumed)	0.88 (assumed)	0.47 (0.23)
R Squared	0.0850	0.0850	0.0850	0.1009	0.1011	0.1015
N	2787	2787	2787	2787	2787	2787
Implied Probability Weighting π(.01) by Experts						
25th Percentile	.	.	.	0.05%	0.11%	1.70%
Median	.	.	.	1.46%	2.35%	11.87%
75th Percentile	.	.	.	5.56%	7.73%	24.73%

Panel B. Estimate of Loss Aversion Based on Local Approximation	**Estimate from Mturk (95% c.i.)**	**Median Forecast (25th, 75th ptile)**
	(1)	(2)
Reference Dependence Parameter		
Loss Aversion λ	1.73 (0.26, 5.08)	2.75 (0.59,8.71)

Notes: Panel A reports the structural estimates of the model in Section 2 using a non-linear least squares regression for observations in the 3 benchmark treatments and in the 2 probabilistic pay treatments. We estimate the model under two assumptions, a power cost of effort function (Columns 1-3) and an exponential cost of effort function (Columns 4-6). The specification reports the estimate for a probability weighting coefficient under the assumption of linear value function (Columns 1 and 4), concave value function with curvature 0.88 as in Tversky and Kahneman (Columns 2 and 5) and with estimated curvature (Columns 3 and 6). Panel B shows the estimates for the loss aversion parameter, which is obtained with a local approximation, see text.

Appendix Figures 1a-d. MTurk Task, Examples of Screenshots

Appendix Figure 1a. Screenshot for 10-cent benchmark treatment, Instructions

On the next page you will play a simple button-pressing task. The object of this task is to alternately press the 'a' and 'b' buttons on your keyboard as quickly as possible for 10 minutes. Every time you successfully press the 'a' and then the 'b' button, you will receive a point. Note that points will only be rewarded when you **alternate** button pushes: just pressing the 'a' or 'b' button without alternating between the two will not result in points.

Buttons must be pressed by hand only (key-bindings or automated button-pushing programs/scripts cannot be used) or task will not be approved.

Feel free to score as many points as you can.

As a bonus, you will be paid an extra 10 cents for every 100 points that you score. This bonus will be paid to your account within 24 hours.

Appendix Figure 1b. Screenshot for 10-cent benchmark treatment, Task

Press 'a' then 'b'...

Points: 302
Bonus Payout: $ 0.30

You will be paid an extra 10 cents for every 100 points that you score.

Appendix Figure 1c. Screenshot for 40-cent gain treatment, Instructions

On the next page you will play a simple button-pressing task. The object of this task is to alternately press the 'a' and 'b' buttons on your keyboard as quickly as possible for 10 minutes. Every time you successfully press the 'a' and then the 'b' button, you will receive a point. Note that points will only be rewarded when you **alternate** button pushes: just pressing the 'a' or 'b' button without alternating between the two will not result in points.

Buttons must be pressed by hand only (key-bindings or automated button-pushing programs/scripts cannot be used) or task will not be approved.

Feel free to score as many points as you can.

As a bonus, you will be paid an extra 40 cents if you score at least 2,000 points. This bonus will be paid to your account within 24 hours.

Appendix Figure 1d. Screenshot for 40-cent gain treatment, Task

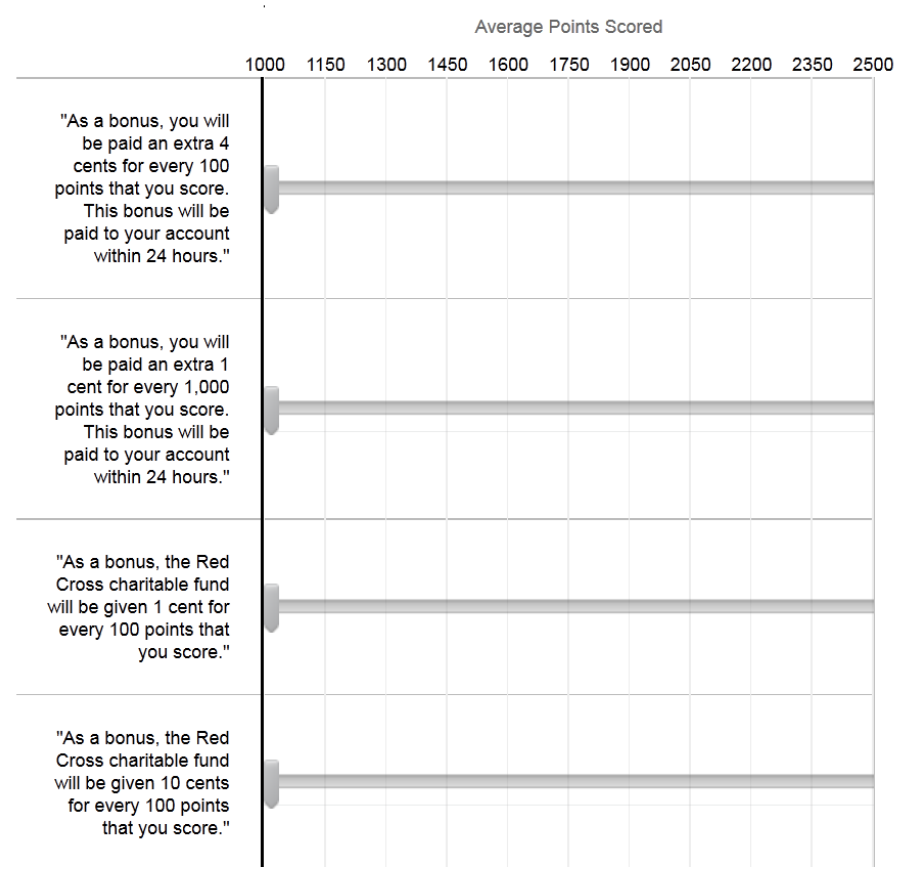
Press 'a' then 'b'...

Points: 215
Bonus Payout: $ 0.00

You will be paid an extra 40 cents if you score at least 2,000 points.

Notes: Appendix Figures 1a-d plot excerpts of the MTurk real-effort task for two treatments, the 10-cent piece rate benchmark treatment (Appendix Figure 1a-b) and the 40-cent gain treatment (Appendix Figure 1c-d). For each treatment, the first screenshot reproduces partially the instructions, while the second screenshot displays the task. These two screens are the only places in which the treatments differed.

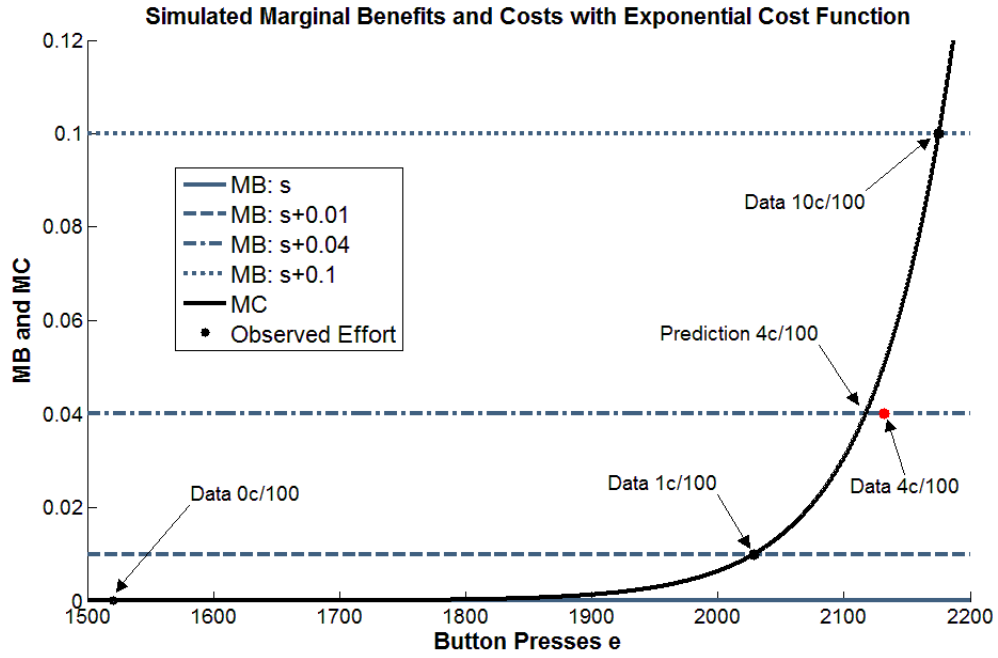
Appendix Figure 2. Expert Survey, Screenshot



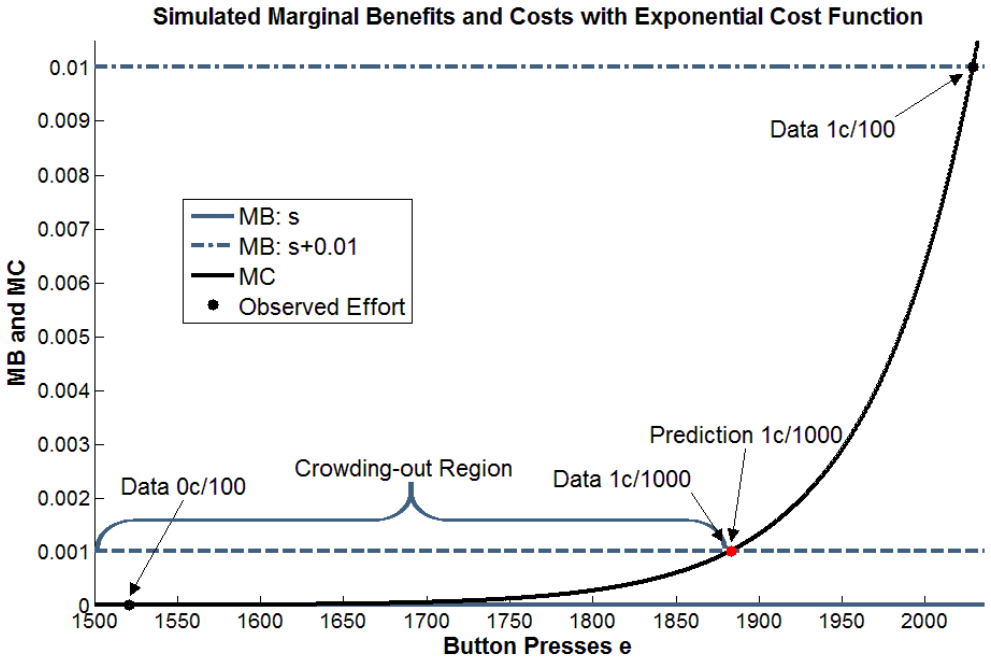
Notes: Appendix Figure 2 shows a screenshot reproducing a portion of the Qualtrics survey which experts used to make forecasts. The survey had 15 sliders, one for each treatment (given that the results for 3 treatments were provided as a benchmark). For each treatment, the left side displays the treatment-specific wording which the subjects assigned to that treatment saw, and on the right side a slider which the experts can move to make a forecast.

Appendix Figure 3. Estimate of Model, Alternative Cost Function (Exponential Cost Function)

Appendix Figure 3a. Estimate with 0c, 1c, 10c Piece Rate, Prediction for 4c Piece Rate (Exponential)

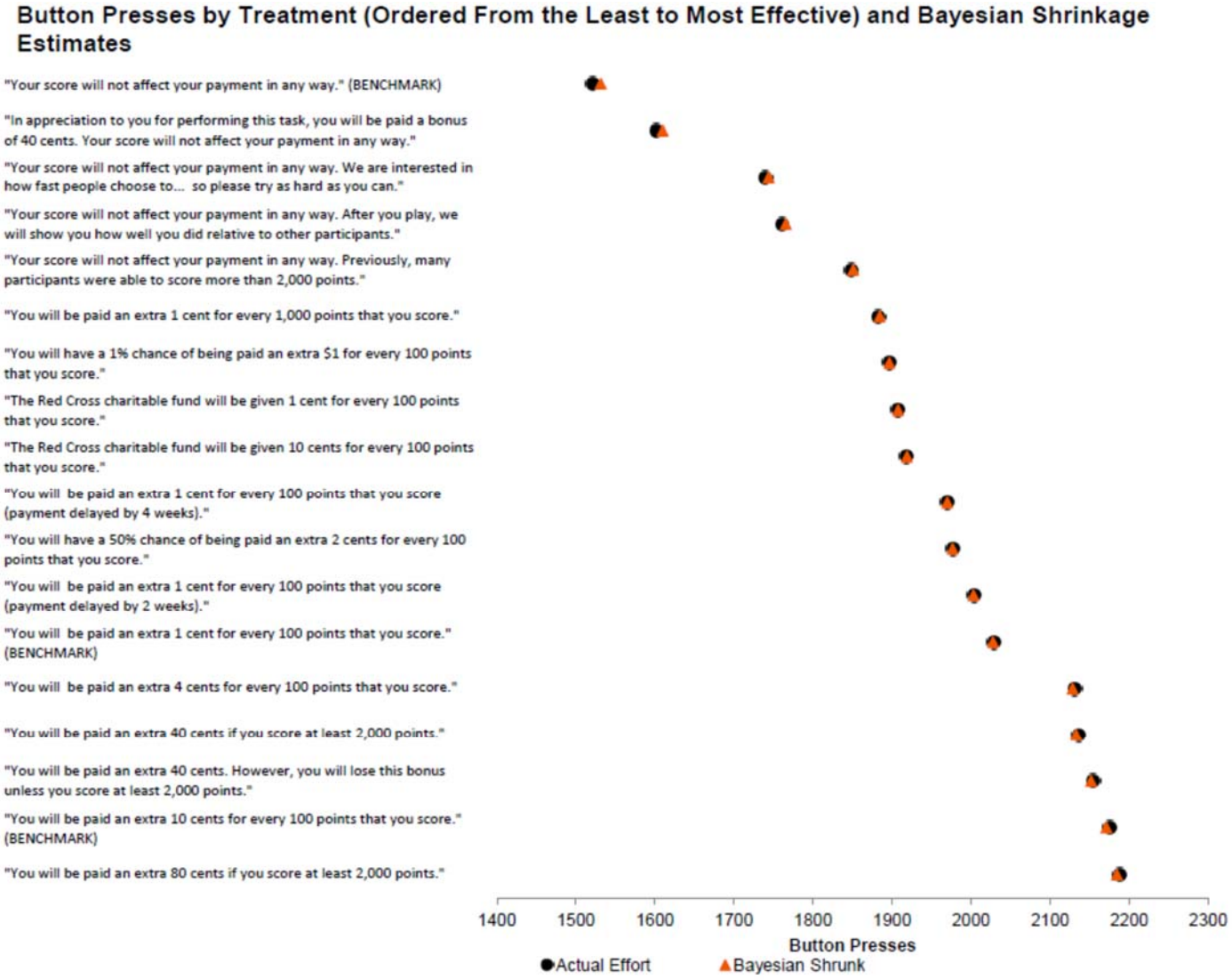


Appendix Figure 3b. Predicted Effort for "Paying Too Little" treatment (Exponential)



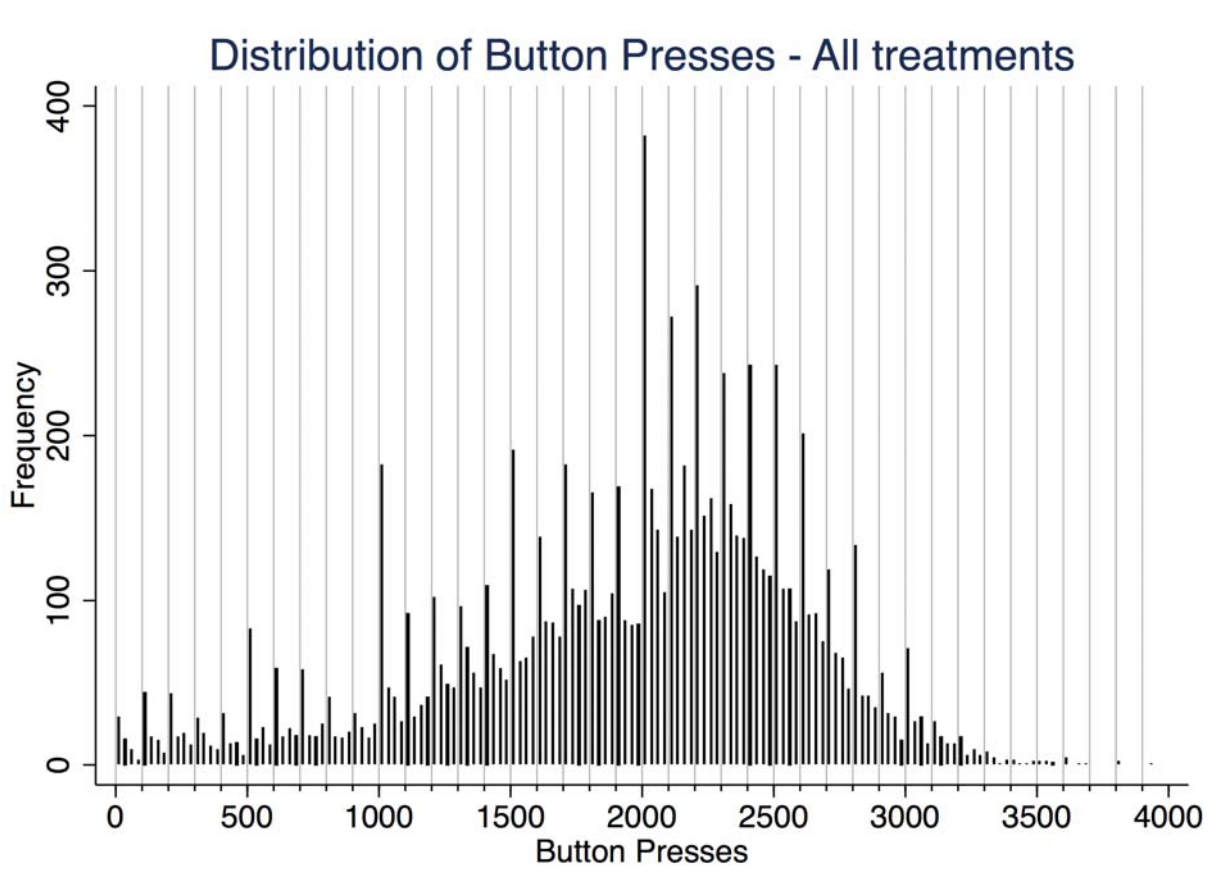
Notes: Appendix Figures 3a-b plot the equivalent of Figures 2a-b, but estimated with an exponential cost function as opposed to a power cost function. Appendix Figure 3a plots the marginal cost curve and the marginal benefit curve for the three benchmark treatments. The figure also plots the out of sample prediction for the 4 cent treatment (which is not used in the estimates), as well as the observed effort for that treatment. Appendix Figure 3b plots, for the same point estimates, the out of sample prediction for the treatment with 1-cent per 1,000 clicks.

Appendix Figure 4. Effort by Treatment, Average and Bayesian Shrinkage Estimator



Notes: Appendix Figure 4 plots the average effort by treatment as in Figure 3, with in addition a Bayesian shrinkage-adjusted measure, to correct for the sampling error (see text for detail). The adjustment makes only a minimal difference.

Appendix Figure 5. Distribution of Button Presses, All Treatments



Notes: Appendix Figure 5 plots a histogram of the observed button presses over all 18 treatments in the real-effort MTurk experiment in bins of 25 points. Notice the spikes at round numbers, in part because incentives kick in at round-number points.

Appendix Table 1. Estimates of Behavioral Parameters, Robustness

Cost of Effort Specification:	**Exponential Cost of Effort**			
Estimation Method:	**Non-linear Least Squares Estimator on Individual Effort**			
Assumption:	*Low Cost Function Curvature*	*High Cost Function Curvature*	*Concave Value Function*	*Continuous Points*
	(1)	(2)	(3)	(4)
Panel A. Estimate of Model on Effort in 3 Benchmark Treatments				
Curvature γ of Cost of Effort Function	0.010 (assumed)	0.020 (assumed)	0.0138 (0.003)	0.0159 (0.0040)
Level *k* of Cost of Effort Function	2.41E-11 (4.46E-12)	1.80E-20 (6.61E-21)	1.34E-14 (9.78E-14)	1.05E-16 (8.92E-16)
Intrinsic Motivation *s* (cent per 100 points)	9.86E-03 (3.59E-03)	2.98E-05 (2.16E-05)	1.67E-03 (3.49E-03)	3.13E-04 (7.63E-04)
Curvature of Utility Over Piece Rate	1 (assumed)	1 (assumed)	0.88 (assumed)	1 (assumed)
R Squared	0.1509	0.1528	0.1532	0.0911
N	1664	1664	1664	1664
Implied Effort, 4-cent Treatment (Actual Effort 2,132)	2123	2112	2087	2117
Implied Effort, Low-pay Treatment (Actual Effort 1,883)	1763	1928	1820	1884
Panel B. Estimates of Social Preferences and Time Preferences				
	Estimate from Mturk (95% c.i.)	**Median Forecast (25th, 75th ptile)**	**Estimate from Mturk (95% c.i.)**	**Median Forecast (25th, 75th ptile)**	**Estimate from Mturk (95% c.i.)**	**Median Forecast (25th, 75th ptile)**	**Estimate from Mturk (95% c.i.)**	**Median Forecast (25th, 75th ptile)**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Social Preferences Parameters								
Pure Altruism Coefficient α	0.007 (-0.033,0.047)	0.094 (0.007,0.338)	0.002 (-0.010,0.014)	0.051 (7.17E-4,0.696)	0.010 (-0.047,0.066)	0.093 (0.004,0.545)	0.003 (-0.017,0.024)	0.067 (0.002,0.538)
Warm Glow Coefficient *a* (cent per 100 points)	0.432 (0.119,0.745)	0.004 (0.000,0.014)	0.060 (-0.027,0.147)	2.61E-05 (-7.60E-5,0.004)	0.510 (-0.030,1.049)	0.001 (0.000,0.013)	0.140 (-0.139,0.419)	2.81E-04 (-3.02E-6,0.007)
Gift Exchange Δ*s* (cent per 100 points)	0.030 (-0.008,0.068)	0.030 (0.005,0.163)	2.99E-04 (-2.0E-4,8.0E-4)	4.62E-04 (3.74E-5,0.009)	0.011 (-0.033,0.055)	0.010 (0.001,0.085)	0.002 (-0.006,0.011)	0.003 (0.000,0.030)
Time Preference Parameters								
Present Bias β	1.74 (-0.53,4.02)	1.31 (0.70,1.72)	0.95 (-1.50,3.40)	0.54 (0.16,0.93)	1.52 (-1.49,4.52)	0.82 (0.34,1.18)	1.15 (-1.29,3.58)	0.76 (0.28,1.16)
(Weekly) Discount Factor δ	0.83 (0.50,1.17)	0.91 (0.75,1.00)	0.70 (0.14,1.25)	0.82 (0.58,1.00)	0.78 (0.31,1.24)	0.87 (0.68,1.00)	0.76 (0.27,1.26)	0.85 (0.65,1.00)

Notes: This table reports the results of four robustness checks, each estimated using a non-linear least squares estimator with an exponential cost of effort function. The specification regresses the effort of the individual MTurker (rounded to the nearest 100 points) with the specification discussed in Section 6. The specification in Panel A include only the 3 benchmark treatments, while the specifications in Panel B include also the charitable giving, gift exchange, and time-delay treatments. For each specification, the first Column in Panel B presents the parameter estimates from the MTurker effort, while the second column presents the implied parameter value for the expert forecast at the median, the 25th percentile and the 75th percentile of the expert distribution. The first two robustness checks examine the impact of mis-specifications in the cost of effort function by forcing the curvature parameter to be fixed at a low value (Column 1) or a high value (Cloumn 2). The second robustness check involves estimates which assume a concave value function, as opposed to linear utility, taking the Tversky and Kahneman 0.88 curvature. Column 4 is like the benchmark, except that, instead of using the points rounded to 100, it uses the continuous points, assuming (for simplicity) that the incentives are distributed continuously.