Real incentives really matter*

Christoph Drobner[†] Sebastian J. Goerg[‡] Orestis Kopsacheilis[§]

July 24, 2026

Abstract

Incentivizing behavior, a core principle of economic experiments, is currently under scrutiny due to a series of papers that find little to no difference in choices made under real or hypothetical incentives. We experimentally assess the effectiveness of hypothetical incentives to induce participants' real effort – an integral aspect of reliable experimental data. We find that although linking behavior to hypothetical incentives generates modest additional effort compared to no reference to such incentives, hypothetical incentives fall substantially short of real ones, even when their nominal value is radically inflated. To understand the mechanism behind these effects, we systematically vary base pay and piece-rate magnitudes across incentive conditions and structurally estimate a model of costly effort. Although higher base pay increases effort both by increasing non-pecuniary motivation and by enhancing responsiveness to hypothetical incentives, incentive-compatible real payments are far more cost-effective.

Keywords: Real Incentives, Hypothetical Incentives, No Incentives, Real Effort

JEL Codes: C90, C91, D01

*This study was pre-registered on AsPredicted and received IRB approval from the German Association for Experimental Economic Research e.V. (Certificate No. rAf5iJta).

†Central European University. Email: drobnerc@ceu.edu.

‡NIM – Nuremberg Institute for Market Decisions and Technical University Munich (TUMCS, SOM). Email: sebastian.goerg@nim.org.

§Technical University Munich (TUMCS, SOM); University of Crete. Email: orestis.kopsacheilis@uoc.gr.

1 Introduction

Incentivizing behavior is a foundational principle of experimental economics. Providing sufficiently high monetary incentives that are linked to the outcome variable has been argued to eliminate other motivations that could bias behavior (Smith, 1982; Plott, 1986; Svorenčík and Maas, 2016). While scholars in other behavioral sciences have long advocated for critical reflection on the use of monetary incentives (see Hertwig and Ortmann 2001 for a relevant discussion), this perspective has only recently gained traction among economists. This is largely due to a recent series of preference-elicitation studies that report little or no difference between measures obtained under incentivized and non-incentivized[1] conditions. For example, the meta-analysis by Matousek et al. (2022) on time preferences concludes that "it does not matter systematically for the reported discount rates whether experiments use real or hypothetical rewards" (p. 320). Moreover, Brañas-Garza et al. (2023) draw on cross-country studies with field and online experiments to demonstrate that time preferences under real and hypothetical monetary incentives are largely comparable. Similarly, Falk et al. (2023) validate a survey module that uses hypothetical monetary incentives, across several preference dimensions. They conclude that these hypothetical measures offer a good compromise and apply this module in one of the largest global preference studies to date (Falk et al., 2018).

A common feature of preference-based choice tasks is that non-incentivized conditions link behavior to monetary rewards even though these are only hypothetical. For instance, Brañas-Garza et al. (2023) elicit choices between smaller, sooner and larger, later hypothetical monetary rewards to measure participants' time preferences. The use of such hypothetical incentives is less common in other behavioral tasks, where non-incentivized conditions typically make no mention of monetary rewards beyond the participation fee. Reviewing a wide range of behavioral tasks, Camerer and Hogarth (1999) find that, unlike in preference-based choice tasks, real monetary incentives affect behavior more systematically in tasks involving judgment, learning, memory and recall, or real-effort provision. The disparity between these two types of non-incentivized conditions raises the question of whether linking behavior to hypothetical incentives differs from not mentioning any link to monetary rewards and how both compare to real ones.

Addressing this research question tests whether people can effectively internalize hypothetical

[1] By 'non-incentivized' we refer to conditions where the outcome variable is not linked to real monetary incentives and where participants typically receive only a fixed participation fee.

incentives, treating them as motivationally similar to real ones – a potential explanation for the observed similarity between real and hypothetical incentives in preference-based choice tasks. To this end, we conduct two complementary studies using pre-registered real-effort experiments based on a decoding task (Sillamaa, 1999a). Real-effort tasks enable a clear comparison between settings in which performance is linked to monetary rewards and settings in which incentives are not mentioned. This comparison is not feasible in preference-based choice tasks where monetary outcomes must be specified by design. To the extent that preference-based choice tasks involve some degree of effort, we would expect to see some trace of this internalization ability in a standard real-effort task. Beyond testing this hypothesis, establishing whether hypothetical incentives can motivate effort has practical implications, potentially offering experimenters a cost-effective alternative to real monetary incentives across a broader range of behavioral tasks.

In Study 1 we compare effort provision across three conditions: real 10-pence piece-rate incentives (Real), hypothetical 10-pence piece-rate incentives (Hypothetical), and no mention of incentives (No-Incentives). In all three conditions, participants received a £1 participation fee that is not linked to their output. Hypothetical is identical to Real, except that participants are explicitly informed, and repeatedly reminded, that the piece rate for solving items is hypothetical and does not affect their final payment. We find that Hypothetical increases effort only marginally relative to No-Incentives, an increase that is not significant at the 5% level (Cohen's $d = 0.16$). Real, by contrast, generates substantially larger and statistically significant effects, relative to both No-Incentives ($d = 1.66$) and Hypothetical ($d = 1.41$).

While large effects are not uncommon when comparing performance-contingent pay to flat compensation, in a comparable design, DellaVigna and Pope (2018) report Cohen's d = 1.0, our effect is on the larger side. The magnitude replicates in Study 2, which finds even larger effects for Real (Cohen's d = 2.3) and weaker effects for Hypothetical (Cohen's d = 0.09).

In Study 2 we introduce four treatments that systematically vary base pay and piece-rate parameters across the three incentive conditions. The four additional treatments allow us to identify the following boundary conditions: First, even a ten-fold increase of the hypothetical piece rate has negligible impact. Second, substantially raising the base pay to £3.20, matching the average total compensation in Study 1's Real, generates a modest but statistically significant increase in effort under No-Incentives. Third, this higher base pay increases responsiveness to Hypothetical:

participants receiving £3.20 upfront with a £1.00 hypothetical piece rate solve significantly more items than their counterparts in No-Incentives with the same high base pay. Fourth, this high base pay hypothetical variant is substantially less cost-effective than the real condition. A real piece rate of only £0.01 under low base pay induces higher effort and requires only one-third of the total cost.

To decompose the treatment effects on real effort provision into interpretable parameters, we develop and estimate a structural model of costly effort. We build on DellaVigna and Pope (2018)'s core framework of costly effort choice, in which individuals maximize utility from monetary piece rates and non-monetary motivation (such as gift-exchange reciprocity), minus a convex cost of effort. We extend this model by incorporating a parameter that captures how nominal piece-rates are internalized when they are hypothetical. The net effect of hypothetical piece-rates on optimal effort provision is the product of this parameter and the nominal piece-rate.

The structural estimates reveal that participants internalize hypothetical piece rates at just above 0.1% at the low base pay. This point estimate declines further when the hypothetical piece rate increases. Within our tested parameter range the net effect on effort remains positive, but this pattern implies a boundary condition: excessively high hypothetical piece rates could reduce net effort if internalization declines faster than the nominal piece rate increases. For example, a £1,000 hypothetical piece rate is likely dismissed as implausible, so internalization fails and further nominal increases backfire. Furthermore, the model-mediated analysis reveals two distinct channels through which raising the base pay increases effort. First, it increases non-pecuniary motivation substantially: participants reciprocate by providing additional effort in a way that is consistent with 'gift exchange' (Gneezy and List, 2006; DellaVigna and Pope, 2018; DellaVigna et al., 2022). Second, it significantly increases participants' responsiveness to hypothetical incentives. However, despite this increase, internalization remains very low, staying below 0.3%.

Both model-free and model-mediated analysis converge on the same conclusion: while linking effort to hypothetical incentives can generate some improvements over not referencing any incentives at all, this strategy remains a poor substitute for real monetary rewards. Although our primary evidence comes from a real-effort setting in which effort and performance are tightly linked, our exploratory analysis indicates why our conclusions extend more broadly. We find that in No-Incentives and Hypothetical, the share of participants who exert only minimal effort, i.e., solving no more than one item, is more than three times higher compared to REAL. This concentration of minimal-effort responses suggests not merely low motivation but, as Rydval and Ortmann (2004) observe in a related context, outright refusal to engage with the task. Furthermore, participants in both NO-INCENTIVES and HYPOTHETICAL conditions exhibit significantly higher error rates on a comprehension question which was administered after the introduction of incentives.[2] In contrast, error rates in the practice round for the decoding task, administered before participants learned their incentive condition, show no systematic differences. Returning to our initial motivation, these patterns provide little support for the hypothesis that participants efficiently internalize hypothetical incentives — even when the task requires only minimal effort. Therefore, the null effects between real and hypothetical incentives in preference-based choice tasks likely have a different explanation.

Although our main result clearly supports the use of real monetary incentives, in many experimental settings, e.g., medical decision-making tasks or high-stakes preference elicitation, implementing incentive-compatible payments is sometimes not possible. In these cases, our results offer practical guidance. Generous participation fees can partially substitute incentive compatible payments, producing more effortful, higher-quality responses even when the incentives related to the task must remain hypothetical.

2 Background on Real-Effort Tasks

We investigate whether individuals internalize hypothetical monetary incentives as motivationally equivalent to real ones, using effort provision as our primary dependent variable. We focus on effort to address this fundamental question because insights related to effort extend beyond tasks explicitly designed to measure it. Even when experimental tasks are not explicitly designed to measure effort, ensuring high-quality data requires participants to engage meaningfully with the task – for example, by reading instructions carefully, considering trade-offs thoughtfully, and performing calculations accurately. This focus on effort is consistent with Smith (1982), who formalizes the principle that monetary incentives must dominate subjective costs, including the cognitive effort required for

[2] Error rates are based on participants' initial responses. Importantly, however, to ensure participants responded to our incentive manipulation rather than misunderstanding it, we required all participants to answer correctly before proceeding to the task.

thinking, calculating, and acting, in order to ensure control over participant preferences.

Two primary methodological approaches dominate experimental research on effort provision: stated-effort and real-effort paradigms (see Charness et al., 2018; Carpenter and Huet-Vaughn, 2019, for reviews). In stated-effort paradigms, participants report intended effort levels, with higher stated effort corresponding to greater monetary rewards but also to increased costs (often affecting other participants' payoffs), as in gift-exchange (Fehr et al., 1993) or tournament (Nalbantian and Schotter, 1997) settings. In real-effort paradigms, participants engage in tangible tasks with outcomes determined by actual performance. Carpenter and Huet-Vaughn (2019) classify real-effort tasks into motor tasks such as counting (Abeler et al., 2011) or moving a slider (Gill and Prowse, 2012), and cognitively demanding tasks such as solving puzzles (Charness and Villeval, 2009) or arithmetic problems (Niederle and Vesterlund, 2007).

We employ a real-effort paradigm based on a decoding task (e.g., Sillamaa, 1999a,b; Lévy-Garboua et al., 2009; Drobner and Orhun, 2025). Recent work by Dutcher et al. (2024) demonstrates that stated-effort and real-effort paradigms can produce similar results when opportunity costs are properly controlled, underscoring that both approaches offer valuable insights for different research questions. Our choice of a real-effort paradigm reflects a specific design requirement: the need to create a clean comparison between conditions where incentives are not mentioned at all versus conditions where they are framed as hypothetical. Stated-effort paradigms and preference-based choice tasks inherently tie decisions to monetary outcomes – whether hypothetical or real – making this distinction not possible.

The remaining features of our experimental design are guided by prior literature examining how varying levels of incentives influence real effort provision. This literature reports mixed findings (Bonner et al., 2000). Specifically, Erkal et al. (2018) and Dorner and Lancsar (2023) summarize that some studies report no impact of monetary incentives on effort provision (e.g., Araujo et al., 2016), others document a positive association (e.g., Jenkins Jr et al., 1998; Abeler et al., 2011; DellaVigna and Pope, 2018; De Quidt et al., 2018; Goerg et al., 2019), while a third group finds a non-monotonic relationship (e.g., Gneezy and Rustichini, 2000; Pokorny, 2008; Ariely et al., 2009).

Variation in piece-rate magnitude is a key moderator of this heterogeneity and central to our investigation. DellaVigna and Pope (2018), for example, report Cohen's $d = 1.0$ when comparing a 10-cent piece rate to no incentives, and $d = 0.7$ for a 1-cent piece rate versus no incentives.

Likewise, in Experiment 3 of De Quidt et al. (2018), real effort differs substantially across piece-rate levels: Cohen's $d = 1.1$ when comparing a 4-cent piece rate to no incentives, and $d = 0.7$ for a 1-cent piece rate versus no incentives (pooling over demand treatments). We extend this work by systematically varying both piece-rate magnitudes and base pay levels across real, hypothetical, and no-incentive conditions. In doing so, we incorporate insights from prior literature identifying a range of additional moderators such as intrinsic motivation, skill, time pressure, and opportunity costs that affect effort provision.

First, we use a real-effort task with low intrinsic motivation. Tasks with high intrinsic motivation can confound the relationship between monetary incentives and effort through complex, often non-monotonic interactions – monetary rewards may crowd out intrinsic motivation, or intrinsic motivation may substitute for extrinsic incentives (Deci et al., 1999). We chose the decoding task, which Carpenter and Huet-Vaughn (2019) characterize as having low intrinsic motivation. Second, the decoding task imposes minimal cognitive skill demands which have been shown to compete with incentives in effort provision tasks (Camerer and Hogarth, 1999; Rydval and Ortmann, 2004). Third, we impose no time constraints. Ariely et al. (2009) report evidence consistent with backfiring due to choking under (time) pressure, finding that increasing incentives from low to very high decreased performance in some of their tasks. The generality of this pattern is, however, an open question: Enke et al. (2023) find that very high stakes do not reduce performance but increase response times across four cognitive-bias tasks. While our decoding task differs from both settings, we adopt a precautionary approach and remove time pressure, ensuring that choking through this channel cannot confound our comparisons. Fourth, we allow participants to stop working at any time, ensuring that opportunity costs translate into credible outside options. Without such options, effort responds weakly to incentives: Dutcher et al. (2024) report Cohen's $d = 0.8$ with an outside option versus $d = 0.3$ without; Goerg et al. (2019) find a similar pattern ($d = 0.7$ vs. $d = 0.1$). This point is further supported by Erkal et al. (2018), Eckartz (2014), and Corgnet et al. (2015).

We close this section with two important clarifications. First, the experimental design choices discussed above follow insights from prior literature to maximize participants' responsiveness to incentive variation. These choices do not, however, guarantee elimination of moderating factors, nor is elimination our goal. For example, though we selected the decoding task for its low intrinsic motivation, Dorner and Lancsar (2023) use a similar task and find evidence of intrinsic motivation.

Nonetheless, our treatment comparisons are designed such that moderating factors remain constant across the conditions being compared, securing causal identification. Second, we do not explicitly measure the extent to which moderators such as opportunity costs (Dutcher et al., 2024) or intrinsic motivation (Dorner and Lancsar, 2023) are present in our setting. Measuring these factors and investigating heterogeneity in effort responses across our three incentive conditions would be an interesting extension for future research.

3 Methods

This section describes the experimental design and the structural modeling approach used to investigate participants' effort provision under different incentive schemes. Section 3.1 outlines the main experimental task and explains how effort was measured, followed by a description of the treatment conditions that varied the type and size of incentives participants received. We then detail the participant recruitment, randomization, and procedural aspects of our investigation. Section 3.2 introduces our structural model of costly effort, which extends DellaVigna and Pope (2018) to incorporate responsiveness to hypothetical incentives, and describes the estimation protocol used to recover the structural parameters.

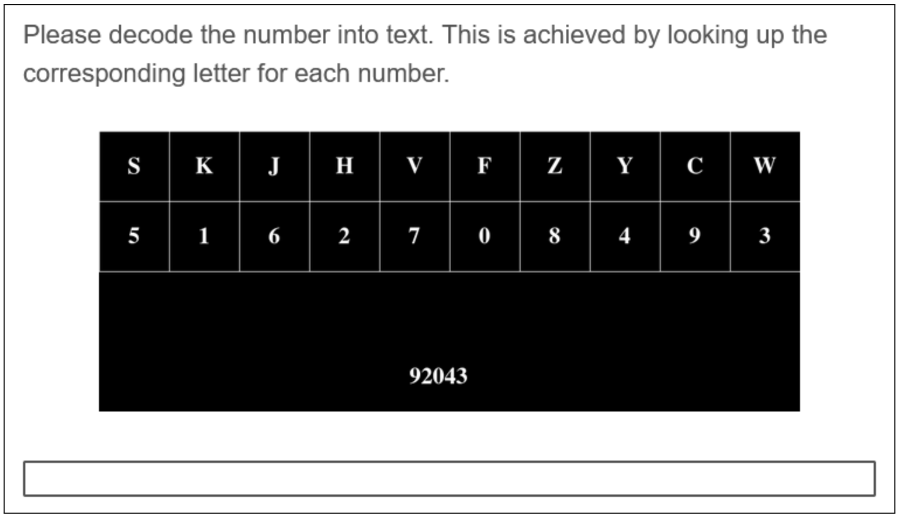
3.1 Experimental Design

We develop an experimental design to test how effort provision responds to real monetary incentives, hypothetical monetary incentives, and no mention of monetary incentives. The primary outcome variable in our study is the number of correctly solved items in a decoding task, which we use as a proxy for effort e. Participants could complete up to 25 decoding items. Each item required translating a five-digit number into text using a provided decoding key, as illustrated in Figure 1. Participants faced no time constraint. All participants were required to complete the first item. After each subsequent item, participants decided whether to continue or stop.[3] Before learning about their specific incentive condition, participants completed a practice round of the decoding task. Besides the number of correctly solved decoding items, our experimental design also tracks

[3]The instructions specified that the participation fee would be paid "regardless of when you decide to stop working" and stopping early carried no penalty beyond forgoing any further piece-rate bonus in the Real treatments: Prolific prohibits rejections based on low task output alone, so participants' base payment and platform reputation were not at risk.

the number of attempted decoding items and time spent on the decoding task. This allows us to consider alternative measures of effort as robustness checks (see Appendix A.3).

Figure 1: Sample Item of Decoding Task



Notes: The figure shows a sample item of the decoding task for which the correct answer is 'CHFYW'. Each item in the decoding task requires translating a 5-digit number into text using a provided decoding key.

We employed a between-subjects design in which participants were assigned to one of seven treatments. Across all treatments, the base pay is a real participation fee that is always paid out. Treatments differ in (i) the magnitude of the base pay, (ii) the magnitude of the piece-rate incentive, and (iii) whether the piece-rate incentive is real or hypothetical. The seven treatments comprise three categories based on incentive type. In Real treatments, participants received a base pay and a piece-rate bonus per correctly solved item that was paid out. In Hypothetical treatments, participants received a base pay and a piece-rate bonus per correctly solved item, but were informed this piece-rate bonus was hypothetical and would not be paid out. In No-Incentives treatments, participants received only a base pay with no mention of bonus payments for solving the decoding task. To mitigate potential selection effects stemming from differences in base payments across treatments, the study was advertised with a uniform participation fee of £1. Participants in treatments with higher base payments were informed of the higher amount during the experiment. Table 1 provides excerpts from the instructions that highlight the treatment differences. Complete experimental instructions are provided in Appendix B.

To ensure that participants fully understood the incentive scheme specific to their treatment,

Table 1: Summary of Treatments

Category	**Treatment**	**Base Pay**	**Piece Rate**	**Description**
No-Incentives	No-B100	£1.00	–	*You will receive the participation fee of **£[Base Pay]** regardless of when you decide to stop working. You will receive **no bonus** payment for solving the decoding tasks.*
	No-B320	£3.20	–	
Hypothetical	Hyp-B100-P10	£1.00	£0.10	*You will receive the participation fee of **£[Base Pay]** regardless of when you decide to stop working. You will receive a bonus payment of **£[Piece-Rate]** for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account.*
	Hyp-B100-P100	£1.00	£1.00	
	Hyp-B320-P100	£3.20	£1.00	
Real	Real-B100-P1	£1.00	£0.01	*You will receive the participation fee of **£[Base Pay]** regardless of when you decide to stop working. You will receive a bonus payment of **£[Piece-Rate]** for each correctly solved decoding task. This bonus payment will be credited to your Prolific account.*
	Real-B100-P10	£1.00	£0.10	

Note: Treatments indicate payment conditions, where "B" denotes the base pay amount and "P" denotes the piece rate.

we incorporated two design features. First, before beginning the task, participants were required to correctly answer a comprehension-check question about both the incentive structure and the nature of the task. Second, after submitting an answer to each item in the decoding task, they were reminded of their incentive scheme before deciding whether to continue.

We collected data for the seven treatments in two pre-registered studies on Prolific, using a sample from the UK. Data were collected in three waves and subjects were randomly assigned to one of the treatments included in the respective wave. The distribution of participants' demographic characteristics is presented in Appendix A.1 and shows that the experimental treatment groups are well balanced. Study 1 was carried out in wave 1 that took place in October 2024. It included the treatments No-B100, Hyp-B100-P10, and Real-B100-P10 (pre-registration). Study 2 was carried out in two waves. Wave 1 took place in August 2025. It included a direct replication of the treatments of Study 1 and additionally introduced the treatments No-B320 and Hyp-B100-P100 (pre-registration). Wave 2 took place one week later. There, we collected data for two additional treatments, Hyp-B320-P100 and Real-B100-P1 (pre-registration). We recruited approximately 300 participants in each treatment within each study, yielding a total sample size of 901 participants in Study 1 and 2,116 participants in Study 2.

3.2 Structural Model

We extend the model of costly effort from DellaVigna and Pope (2018) to incorporate responsiveness to hypothetical incentives. Individuals choose effort level e to maximize utility net of effort costs. For each unit of effort, individuals receive utility from a monetary piece rate p as well as a non-monetary component $s > 0$. The former represents standard performance incentives, while the latter captures intrinsic motivation, sense of duty, or gratitude for the participation fee.

We assume a convex cost of effort function $c(e)$ with $c'(e) > 0$ and $c''(e) > 0$ for all $e > 0$. Individuals solve:

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

We adopt the power cost function $c(e) = ke^{1+\gamma}/(1+\gamma)$, which features a constant elasticity of effort $1/\gamma$ with respect to the marginal benefit of effort. This yields the first-order condition:

$$c'(e) = ke^{\gamma} = s + p, \tag{2}$$

with optimal effort:

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

We extend this framework to allow for partial internalization of hypothetical incentives, i.e., the extent to which participants treat hypothetical rewards as motivating real effort. Let I_{Hyp} and I_{Real} denote indicator variables for hypothetical and real incentive treatments, respectively. The extended first-order condition becomes:

$$c'(e) = ke^{\gamma} = s + \beta \cdot p \cdot I_{\text{Hyp}} + p \cdot I_{\text{Real}}. \tag{4}$$

The parameter β captures the efficiency of hypothetical incentives relative to real incentives. We impose no restrictions on β during estimation: values in $[0, 1]$ indicate partial internalization (with $\beta = 0$ representing complete disregard and $\beta = 1$ representing full equivalence with real incentives), while $\beta > 1$ would indicate hypothetical incentives are more effective than real ones, and $\beta < 0$ would indicate a backfiring effect where hypothetical rewards reduce effort below baseline.

When $I_{\text{Hyp}} = 0$, equation (4) reduces to the standard model. When $I_{\text{Hyp}} = 1$, the hypothetical piece rate contributes βp to the marginal benefit, allowing us to estimate the efficiency of hypothetical incentives.

Following DellaVigna and Pope (2018), we recover the structural parameters (k, γ, s) using average observed effort $\bar{e}$ in three treatments with varying real piece rates but equal participation fee (£1.00): No-B100 ($p = 0$), Real-B100-P1 ($p = 0.01$), and Real-B100-P10 ($p = 0.10$). From equation (4) with $I_{\text{Hyp}} = 0$ and $I_{\text{Real}} = 1$, these three moments exactly identify the structural parameters via the first-order condition $ke^{\gamma} = s + p$. Given convexity, this system has a unique solution, which we solve numerically. We denote the baseline non-pecuniary motivation as s_L to distinguish it from the elevated motivation s_H under high participation fees, yielding estimates $(\hat{k}, \hat{\gamma}, \hat{s}_L)$. Confidence intervals for all structural parameters are derived via bootstrap with 1,000 draws, resampling participants within each treatment.

Given the core parameter estimates, we calculate the responsiveness to hypothetical incentives and the effect of increased participation fees. Rearranging equation (4) under hypothetical incentives ($I_{\text{Hyp}} = 1$, $I_{\text{Real}} = 0$) yields $\beta = (ke^{\gamma} - s)/p$. For treatments with low participation fee (£1.00), we substitute the core estimates $(\hat{k}, \hat{\gamma}, \hat{s}_L)$ and $\bar{e}$ from Hyp-B100-P10 (piece rate £0.10) and Hyp-B100-P100 (piece rate £1.00) to recover $\hat{\beta}_{LL}$ and $\hat{\beta}_{LH}$.[4] We impose no restrictions on β, allowing for values below zero (backfiring effects) or above one (hypothetical incentives more effective than real).

Next, we isolate the effect of increasing the participation fee from £1.00 to £3.20 using No-B320, which features the higher participation fee but no piece rate. With $p = 0$, equation (4) simplifies to $s_H = ke^{\gamma}$. Using the core parameter estimates $(\hat{k}, \hat{\gamma})$ and $\bar{e}$ from No-B320, we recover $\hat{s}_H$. Following DellaVigna and Pope (2018), we interpret the increase $\Delta s = s_H - s_L$ as a gift-exchange effect: participants reciprocate the (unexpected) participation fee increase through higher effort.

[4] Subscripts on β denote (base pay level, hypothetical piece rate level): L indicates low (£1.00 base pay or £0.10 piece rate) and H indicates high (£3.20 base pay or £1.00 piece rate).

Finally, for Hyp-B320-P100, which combines high participation fee (£3.20) and high hypothetical piece rate (£1.00), we rearrange equation (4) to yield $\beta = (ke^{\gamma} - s)/p$. We substitute the core estimates $(\hat{k}, \hat{\gamma})$, the higher motivation parameter $\hat{s}_H$, and $\bar{e}$ from Hyp-B320-P100 to recover $\hat{\beta}_{HH}$. This allows us to test whether higher non-pecuniary motivation enhances responsiveness to hypothetical incentives.

4 Results

We organize the results into two parts. In Section 4.1, we report all pre-registered treatment comparisons and additional exploratory analyses, without imposing any theoretical framework.[5] In Section 4.2, we present the structural estimation results from our model of costly effort.

4.1 Treatment Effects

Figure 2 presents decoding scores, i.e., the number of correctly solved items, across all treatments.[6] Bars display mean scores with their 95% confidence intervals from Study 2. Individual observations are overlaid as semi-transparent points to show the underlying distributions; diamonds indicate Study 1 means.

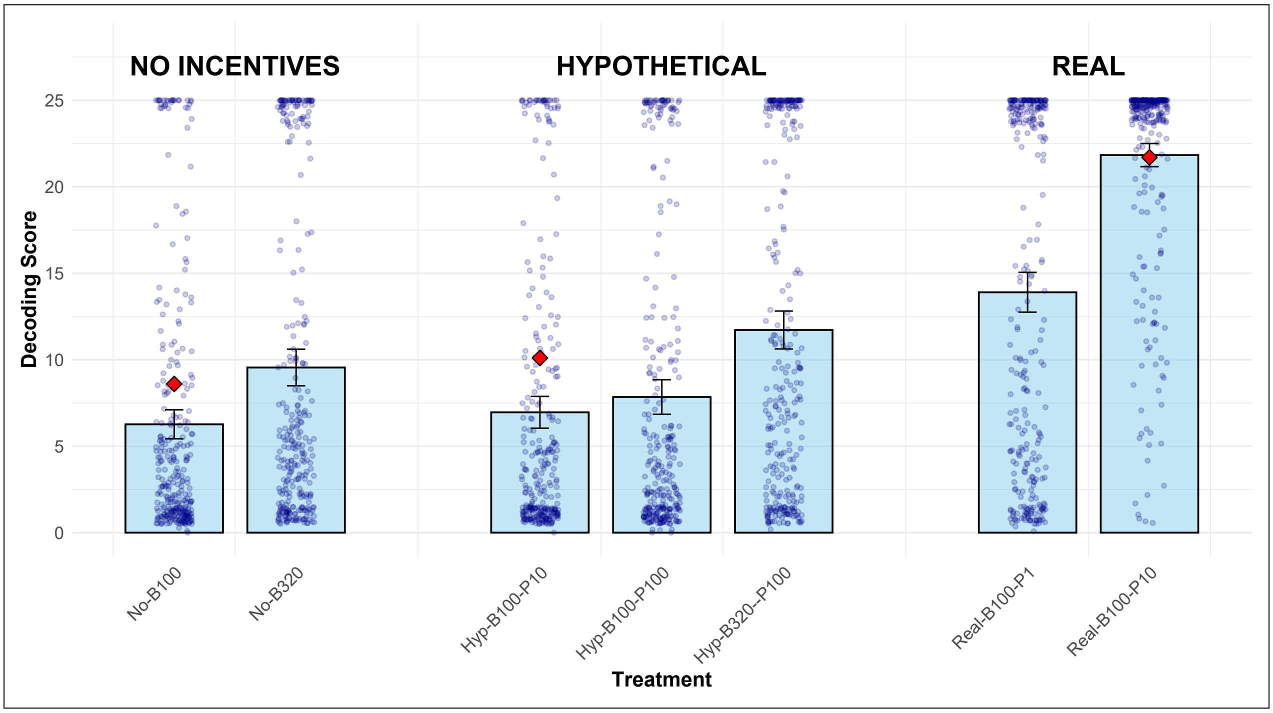
Unless stated otherwise, reported p-values derive from two-sided Wilcoxon rank-sum tests. In Study 1, real incentives generate substantially higher effort: participants in Real-B100-P10 solve 21.8 items on average, compared to 8.6 in No-B100 and 10.1 in Hyp-B100-P10, more than doubling output in both cases (Cohen's $d = 1.66$ and $d = 1.41$; both $p < 0.001$). Hypothetical incentives yield a small increase over the no-incentives baseline ($d = 0.16$, $p = 0.091$), significant only at the 10% level.

In Study 2, we replicate these findings and introduce four additional treatments. Participants in Real-B100-P10 again achieve the highest decoding scores, more than tripling output relative to No-B100 (21.8 vs 6.3; Cohen's $d = 2.30$) and Hyp-B100-P10 (21.8 vs 7.0; Cohen's $d = 2.09$; both $p < 0.001$). Hypothetical incentives again yield minimal effects: scores in Hyp-B100-P10 exceed No-B100 by only 0.7 items (Cohen's $d = 0.09$, $p = 0.747$), an effect smaller than in Study 1 and

[5] Appendix A.2 provides a summary table of pre-registered treatment comparisons.

[6] Appendix A.3 provides cumulative distribution functions for alternative effort measures (i.e., the number of attempted items or time spent on the decoding task) and demonstrates the robustness of our reported results.

Figure 2: Decoding Scores by Treatment



Notes: Bars display mean scores with their 95% confidence intervals from Study 2. Individual observations are overlaid as semi-transparent points to show the underlying distributions; diamonds indicate Study 1 means.

not significant even at the 10% level.

Next, we test whether effort in non-incentivized conditions can be increased by raising hypothetical piece rates, raising base pay, or combining both. Raising the hypothetical piece rate tenfold from £0.10 to £1.00 produces minimal change: Hyp-B100-P100 yields 7.8 items versus 7.0 in Hyp-B100-P10 (Cohen's $d = 0.10$, $p = 0.256$). Increasing base pay alone from £1.00 to £3.20 modestly increases effort: No-B320 yields 9.6 items versus 6.3 in No-B100 (Cohen's $d = 0.39$, $p < 0.001$). The best-performing hypothetical condition combines high base pay with high piece rates: Hyp-B320-P100 yields 11.7 items versus 9.6 in No-B320 (Cohen's $d = 0.23$, $p = 0.005$). Yet in an exploratory analysis we find that even this combination underperforms a real £0.01 piece rate at a fraction of the cost: Real-B100-P1 yields 13.9 items at an average total cost of approximately £1.14 per participant, versus 11.7 items at £3.20 in Hyp-B320-P100, though this difference is significant only at the 10% level (Cohen's $d = 0.22$, $p = 0.053$).

Although Study 2 confirms Study 1's core finding, that real monetary incentives elicit two to three times more effort than any other condition, performance in No-B100 and Hyp-B100-P10 declined between Study 1 (October 2024) and Study 2 (August 2025), whereas Real-B100-P10 remained remarkably stable (Figure 2). This divergence accounts for the larger treatment effects observed in Study 2 compared to Study 1.

Looking at the distribution of individual observations in Figure 2 reveals further differences across treatment categories. In No-Incentives and Hypothetical conditions, nearly one-third of participants solve at most one item (29.4% and 31.2% respectively), compared to only 9.6% in Real (χ^2-test vs. Real: both $p < 0.001$). The pattern reverses at maximum effort: 56.3% of Real participants solve at least 24 items, compared to 16.1% and 19.8% in No-Incentives and Hypothetical (χ^2-test vs. Real: both $p < 0.001$).

Before proceeding to the decoding task, participants were required to (a) solve a practice decoding item and (b) correctly answer a comprehension check question. Error rates in the practice round for the decoding task, administered before participants learned their incentive condition, show no systematic differences: 5.3% in No-Incentives, 5.1% in Hypothetical, and 5.3% in Real (all pairwise $p > 0.46$).

In contrast, error rates in the comprehension check question, administered after participants learned their incentive condition, vary significantly across treatment categories 33.4% in No-Incentives, 28.7% in Hypothetical, and 18.6% in Real (χ^2-test: $p < 0.001$ for Real vs both other categories; Hypothetical vs No-Incentives, $p = 0.052$).[7] In Appendix A.4.2, we unpack the comprehension check error rates by treatment and by incentive-related versus task-related items. The analysis shows that the effect is largely driven by errors on incentive-related items: 29.3% in No-Incentives, 22.7% in Hypothetical, and 11.6% in Real (χ^2-test: $p < 0.001$ for Real vs. both other categories; Hypothetical vs. No-Incentives, $p = 0.004$).

4.2 Structural Estimation

Table 2 presents the structural parameter estimates. The cost function exhibits moderate curvature ($\hat{\gamma} = 5.06$), yielding an elasticity of effort with respect to the piece rate of approximately $1/\hat{\gamma} = 0.20$. A tenfold increase in piece rates thus raises effort by approximately 58.5% ($10^{0.20} = 1.585$), consistent with the observed increase when raising the real piece rate from £0.01 to £0.10 under

[7] Participants had to answer the comprehension question correctly before starting the decoding task. Consequently, Appendix A.4.1 confirms that our main results remain unchanged when controlling for comprehension errors.

low base pay. This remains well below the proportional response that would occur under linear costs. However, the elasticity in our decoding task is notably higher than DellaVigna and Pope (2018)'s estimate of $1/\gamma = 0.03$ in a button-pressing task, where a tenfold piece-rate increase raised effort by less than 10%.

At low base pay and a £0.10 hypothetical piece rate, participants internalize approximately 0.13% of the nominal incentive ($\hat{\beta}_{LL}$). Increasing the hypothetical piece rate to £1.00 reduces the point estimate of this rate to roughly one-third, approximately 0.04% ($\hat{\beta}_{LH}$). However, the confidence interval for $\hat{\beta}_{LL}$ includes zero while that for $\hat{\beta}_{LH}$ does not, indicating greater variance in responses to the lower hypothetical piece rate. Because the tenfold increase in p outweighs the threefold decline in β, the product $\beta \cdot p$ still increases, so higher hypothetical piece rates raise effort within our tested range of p.

Increasing base pay from £1.00 to £3.20 raises non-pecuniary motivation from $\hat{s}_L = 1.80 \times 10^{-4}$ to $\hat{s}_H = 1.52 \times 10^{-3}$, an approximately eightfold increase. For comparison, DellaVigna and Pope (2018) estimate a roughly sixfold increase in non-pecuniary motivation. However, this disparity could be explained by the fact that in their case, participants experienced an increase in base pay of 40 cents compared to £2.20 in our case. Higher base pay also raises responsiveness to hypothetical incentives. Holding the hypothetical piece rate fixed at £1.00, the internalization parameter increases from $\hat{\beta}_{LH} = 3.81 \times 10^{-4}$ to $\hat{\beta}_{HH} = 2.77 \times 10^{-3}$, an approximately sevenfold increase. Thus, higher base pay increases both baseline effort and participants' responsiveness to hypothetical piece rates.

Appendix Figure A5 illustrates the structural model by inverting it to plot predicted effort as a function of total per-participant budget under real piece-rate incentives, hypothetical piece-rate incentives, and base pay only. This counterfactual analysis is subject to linearity assumptions on how non-pecuniary motivation and internalization vary with base pay.

5 Conclusion

A core tenet of experimental economics holds that behavior should be sufficiently incentivized, yet recent evidence suggests that real and hypothetical incentives may yield similar outcomes in certain contexts such as choice-based preference tasks. In this paper we test whether framing incentives

Table 2: Structural Parameter Estimates

	Estimate	95% CI
Cost function parameters		
Curvature $\hat{\gamma}$	5.06	(4.10, 6.25)
Level $\hat{k}$	1.65×10^{-8}	$(4.59 \times 10^{-10}, 3.10 \times 10^{-7})$
Non-pecuniary motivation		
$\hat{s}_L$ (£1.00 base pay)	1.80×10^{-4}	$(3.43 \times 10^{-5}, 6.98 \times 10^{-4})$
$\hat{s}_H$ (£3.20 base pay)	1.52×10^{-3}	$(4.65 \times 10^{-4}, 3.86 \times 10^{-3})$
$\Delta\hat{s} = \hat{s}_H - \hat{s}_L$	1.34×10^{-3}	—
Responsiveness to hypothetical incentives		
$\hat{\beta}_{LL}$ (£1.00 base pay, £0.10 hyp. piece-rate)	1.26×10^{-3}	$(-1.01 \times 10^{-3}, 5.05 \times 10^{-3})$
$\hat{\beta}_{LH}$ (£1.00 base pay, £1.00 hyp. piece-rate)	3.81×10^{-4}	$(3.68 \times 10^{-5}, 1.20 \times 10^{-3})$
$\hat{\beta}_{HH}$ (£3.20 base pay, £1.00 hyp. piece-rate)	2.77×10^{-3}	$(6.37 \times 10^{-4}, 6.24 \times 10^{-3})$

Notes: Cost function parameter estimates $\hat{k}$, $\hat{\gamma}$, and non-pecuniary motivation estimate $\hat{s}_L$ recovered from three moments – average effort in No-B100, Real-B100-P1, and Real-B100-P10 – under a power cost function. $\hat{s}_H$ calculated from treatment mean No-B320 using $\hat{k}$ and $\hat{\gamma}$. $\hat{\beta}_{LL}$ and $\hat{\beta}_{LH}$ calculated from treatment means Hyp-B100-P10 and Hyp-B100-P100 using $\hat{k}$, $\hat{\gamma}$, and $\hat{s}_L$. $\hat{\beta}_{HH}$ calculated from treatment mean Hyp-B320-P100 using $\hat{k}$, $\hat{\gamma}$, and $\hat{s}_H$. Confidence intervals derived via bootstrap with 1,000 draws, then taking 2.5th and 97.5th percentiles. $\Delta\hat{s}$ computed from point estimates.

in hypothetical terms can stimulate real effort. We conduct two pre-registered studies using a standard real-effort task. Study 1 compares three incentive structures: real monetary incentives, hypothetical monetary incentives, and no mention of incentives. Study 2 replicates Study 1 and introduces additional treatments that systematically varied base pay and piece-rate magnitude within these incentive structures.

We find that although hypothetical incentives can generate some additional effort compared to providing no mention of incentives, they fall substantially short of the effort induced by real incentives, even when their nominal value is radically inflated. In the absence of real piece rates, substantially raising the base pay increases effort modestly, especially when combined with high hypothetical piece rates. However, this strategy remains cost-ineffective as even minimal real piece rates achieve significantly higher effort at a fraction of the total cost.

To decompose these treatment effects into interpretable parameters, we extend a model of costly effort, incorporating a parameter that captures responsiveness to hypothetical incentives, and estimate it structurally. Participants internalize hypothetical piece rates at extremely low rates, with evidence suggesting this rate declines further as hypothetical piece rates increase. Raising base pay substantially increases non-pecuniary motivation, consistent with gift-exchange reciprocity toward the experimenter. Higher base pay also amplifies responsiveness to hypothetical incentives. This analysis identifies and qualitatively characterizes interactions between base pay, hypothetical piece-rate magnitude, and the rate at which participants internalize hypothetical incentives. Measuring these interactions quantitatively would require pinpointing the functional forms that govern them, which remains an open question. As a first step, we impose linearity on how non-pecuniary motivation and internalization vary with base pay and trace out the resulting budget-constrained counterfactuals (Appendix Figure A5), which further illustrate that real piece-rate incentives are more cost-effective than increases in base pay. Because these functional forms are assumed rather than identified, fully answering the question would require additional treatments that introduce further variation in piece rates and base pay levels.

Another important question concerns the extent to which our findings generalize beyond real-effort tasks. Two features of our results speak to this. First, the stark increase in effort induced by real monetary incentives replicated consistently across both studies. This combination of magnitude and robustness constitutes a necessary condition for generalizing a finding beyond the immediate experimental context. Second, even though our findings come from a real-effort task, when effort is broadly defined to account for the time and cognitive resources that participants invest, our results become relevant across task types.

One possible objection is that preference-based choice tasks might require significantly less effort than ours; though whether they do is itself an empirical question. More importantly, it is worth asking whether there exists a minimum effort threshold below which data quality deteriorates in such tasks. Some features of our exploratory analysis may offer a useful starting point for this investigation: our results suggest that when real incentives were absent, a substantial share of participants exerted minimal effort, solving at most one item, and showed higher error rates on the comprehension check. However, since both analyses are exploratory, these patterns require further investigation. Future research could systematically assess the minimum level of effort different task types require to yield high-quality data (including instruction comprehension, genuine introspection, and accurate calculations) and examine whether hypothetical incentives can reliably sustain this baseline engagement. Thus, while we refrain from drawing strong conclusions about why some preference-based choice studies find null effects between hypothetical and real incentives, our results provide clear evidence that when experimental tasks impose real costs, real incentives matter.

Statements and Declarations

Conflict of interest: The authors have no conflict of interest.

Ethics approval: This study received IRB approval from the German Association for Experimental Economic Research e.V. (Certificate No. rAf5iJta).

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2):470–492.

Araujo, F. A., Carbone, E., Conell-Price, L., Dunietz, M. W., Jaroszewicz, A., Landsman, R., Lamé, D., Vesterlund, L., Wang, S. W., and Wilson, A. J. (2016). The slider task: An example of restricted inference on incentive effects. *Journal of the Economic Science Association*, 2:1–12.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, 76(2):451–469.

Bonner, S. E., Hastie, R., Sprinkle, G. B., and Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of management accounting research*, 12(1):19–64.

Brañas-Garza, P., Jorrat, D., Espín, A. M., and Sánchez, A. (2023). Paid and hypothetical time preferences are the same: Lab, field and online evidence. *Experimental Economics*, 26(2):412–434.

Camerer, C. F. and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19:7–42.

Carpenter, J. and Huet-Vaughn, E. (2019). Real-effort tasks. In *Handbook of research methods and applications in experimental economics*, pages 368–383. Edward Elgar Publishing.

Charness, G., Gneezy, U., and Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149:74–87.

Charness, G. and Villeval, M.-C. (2009). Cooperation and competition in intergenerational experiments in the field and the laboratory. *American Economic Review*, 99(3):956–978.

Corgnet, B., Hernán-González, R., and Schniter, E. (2015). Why real leisure really matters: Incentive effects on real effort in the laboratory. *Experimental Economics*, 18:284–301.

De Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6):627.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2022). Estimating social preferences and gift exchange at work. *American Economic Review*, 112(3):1038–1074.

DellaVigna, S. and Pope, D. (2018). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

Dorner, Z. and Lancsar, E. (2023). Don't pay the highly motivated too much. *Journal of Behavioral and Experimental Economics*, 103:101972.

Drobner, C. and Orhun, A. Y. (2025). Misguided effort. *Available at SSRN 5311074*.

Dutcher, E. G., Salmon, T. C., and Saral, K. J. (2024). Is "real" effort more real? *Experimental Economics*, pages 1–31.

Eckartz, K. (2014). Task enjoyment and opportunity costs in the lab: The effect of financial incentives on performance in real effort tasks. Technical report, Jena Economic Research Papers.

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., and Van De Ven, J. (2023). Cognitive biases: Mistakes or missing stakes? *Review of Economics and Statistics*, 105(4):818–832.

Erkal, N., Gangadharan, L., and Koh, B. H. (2018). Monetary and non-monetary incentives in real-effort tournaments. *European Economic Review*, 101:528–545.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Falk, A., Becker, A., Dohmen, T., Huffman, D., and Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science*, 69(4):1935–1950.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent market clearing? an experimental investigation. *The Quarterly Journal of Economics*, 108(2):437–459.

Gill, D. and Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102(1):469–503.

Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.

Gneezy, U. and Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810.

Goerg, S. J., Kube, S., and Radbruch, J. (2019). The effectiveness of incentive schemes in the presence of implicit effort costs. *Management Science*, 65(9):4063–4078.

Hertwig, R. and Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24(3):383–403.

Jenkins Jr, G. D., Mitra, A., Gupta, N., and Shaw, J. D. (1998). Are financial incentives related to performance? a meta-analytic review of empirical research. *Journal of Applied Psychology*, 83(5):777.

Lévy-Garboua, L., Masclet, D., and Montmarquette, C. (2009). A behavioral laffer curve: Emergence of a social norm of fairness in a real effort experiment. *Journal of Economic Psychology*, 30(2):147–161.

Matousek, J., Havranek, T., and Irsova, Z. (2022). Individual discount rates: a meta-analysis of experimental evidence. *Experimental Economics*, 25(1):318–358.

Nalbantian, H. R. and Schotter, A. (1997). Productivity under group incentives: An experimental study. *American Economic Review*, pages 314–341.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Plott, C. R. (1986). Rational choice in experimental markets. *Journal of Business*, pages S301–S327.

Pokorny, K. (2008). Pay—but do not pay too much: An experimental study on the impact of incentives. *Journal of Economic Behavior & Organization*, 66(2):251–264.

Rydval, O. and Ortmann, A. (2004). How financial incentives and cognitive abilities affect task performance in laboratory settings: An illustration. *Economics Letters*, 85(3):315–320.

Sillamaa, M.-A. (1999a). How work effort responds to wage taxation: A non-linear versus a linear tax experiment. *Journal of Economic Behavior & Organization*, 39(2):219–233.

Sillamaa, M.-A. (1999b). How work effort responds to wage taxation: An experimental test of a zero top marginal tax rate. *Journal of Public Economics*, 73(1):125–134.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *American Economic Review*, 72(5):923–955.

Svorenčík, A. and Maas, H. (2016). *The Making of Experimental Economics*. Springer.

A Additional Analyses

A.1 Balance of Treatment Groups

The following analysis demonstrates that our treatment groups are balanced in terms of demographics. Participants provided basic demographic information at the end of the survey. Specifically, we collected data on age, gender identity, education level, and household income.

Age was measured using a categorical item with seven response options: *Under 18, 18–24, 25–34, 35–44, 45–54, 55–64*, and *65+ years old*. Gender identity was measured by asking participants how they describe themselves, with the following response options: *Male, Female, Non-binary/third gender, Prefer to self-describe (PSD)*, and *Prefer not to say (PNS)*. Education level was measured as the highest level of education completed, with the following response options: *Some Primary, Completed Primary School, Some Secondary, Completed Secondary School, Vocational or Similar, Some University but no degree, University Bachelor's Degree, Graduate or Professional Degree*, and *Prefer not to say*. Household income was measured as total pre-tax income during the past 12 months, with the following response options: *Less than £20,000; £20,000–39,999; £40,000–59,999; £60,000–99,999*; and *More than £100,000*.

Table A1 presents the distribution of demographic characteristics across treatment groups in Study 1. Each panel reports percentages and the corresponding p-value from a chi-squared test of independence, which assesses whether the distribution of each variable differs significantly across treatments. All chi-squared tests yield p-values above the conventional 5% significance threshold (Age: $p = 0.826$; Gender: $p = 0.551$; Education: $p = 0.611$; Income: $p = 0.067$), showing no statistically significant differences between treatment groups.

Similarly, Table A2 presents the distribution of demographic characteristics across treatment groups in Study 2. Each panel again reports percentages and the corresponding p-value from a chi-squared test of independence. All chi-squared tests yield p-values above the conventional 5% significance threshold (Age: $p = 0.294$; Gender: $p = 0.791$; Education: $p = 0.130$; Income: $p = 0.070$), showing no statistically significant differences between treatment groups.

Table A1: Balance Table - Study 1

Panel A: Age

	<24	25–34	35–44	45–54	55–64	65+
No-B100	9.36%	30.10%	26.09%	18.39%	9.70%	6.35%
Hyp-B100-P10	9.60%	25.83%	26.16%	22.19%	10.93%	5.30%
Real-B100-P10	10.33%	29.33%	23.67%	17.33%	13.33%	6.00%
Total	9.77%	28.41%	25.31%	19.31%	11.32%	5.88%

p-value (chi-squared test) = 0.826

Panel B: Gender

	Male	Female	Nonbinary	PSD	PNS
No-B100	50.17%	48.83%	0.33%	0.00%	0.67%
Hyp-B100-P10	43.05%	54.64%	0.99%	0.33%	0.99%
Real-B100-P10	44.00%	55.00%	0.33%	0.33%	0.33%
Total	45.73%	52.83%	0.55%	0.22%	0.67%

p-value (chi-squared test) = 0.551

Panel C: Education Level

	C.Prim	S.Sec	C.Sec	Voc	S.Univ	C.Univ	Grad/Prof	PNS
No-B100	0.00%	1.00%	19.06%	14.72%	9.70%	36.12%	18.73%	0.67%
Hyp-B100-P10	0.66%	1.32%	18.87%	13.91%	5.96%	41.39%	17.88%	0.00%
Real-B100-P10	0.33%	1.00%	23.67%	13.67%	9.33%	35.33%	16.00%	0.67%
Total	0.33%	1.11%	20.53%	14.10%	8.32%	37.62%	17.54%	0.44%

p-value (chi-squared test) = 0.611

Panel D: Household Income

	< 20,000	20,000–39,999	40,000–59,999	60,000–99,999	> 100,000
No-B100	16.39%	29.77%	25.08%	22.41%	6.35%
Hyp-B100-P10	8.28%	31.13%	26.16%	25.17%	9.27%
Real-B100-P10	15.33%	34.67%	21.33%	22.33%	6.33%
Total	13.32%	31.85%	24.20%	23.31%	7.33%

p-value (chi-squared test) = 0.067

Table A2: Balance Table - Study 2

Panel A: Age

	<24	25–34	35–44	45–54	55–64	65+
No-B100	4.87%	28.25%	25.65%	19.81%	14.29%	7.14%
No-B320	9.33%	24.67%	26.33%	18.67%	13.67%	7.33%
Hyp-B100-P10	6.67%	24.00%	25.67%	20.00%	14.67%	9.00%
Hyp-B100-P100	5.65%	26.25%	24.92%	19.27%	18.27%	5.65%
Hyp-B320-P100	8.20%	28.85%	22.95%	17.70%	16.72%	5.57%
Real-B100-P1	11.59%	25.50%	21.85%	16.56%	16.89%	7.62%
Real-B100-P10	8.00%	23.00%	25.67%	22.00%	18.00%	3.33%
Total	7.75%	25.80%	24.72%	19.14%	16.07%	6.52%

p-value (chi-squared test) = 0.294

Panel B: Gender

	Male	Female	Nonbinary	PSD	PNS
No-B100	46.10%	51.62%	1.62%	0.32%	0.32%
No-B320	44.67%	54.33%	0.33%	0.33%	0.33%
Hyp-B100-P10	52.00%	47.00%	0.67%	0.00%	0.33%
Hyp-B100-P100	49.50%	49.17%	1.00%	0.00%	0.33%
Hyp-B320-P100	47.21%	51.80%	0.66%	0.33%	0.00%
Real-B100-P1	49.67%	49.67%	0.66%	0.00%	0.00%
Real-B100-P10	48.67%	50.67%	0.00%	0.00%	0.67%
Total	48.25%	50.61%	0.71%	0.14%	0.28%

p-value (chi-squared test) = 0.791

Panel C: Education Level

	C.Prim	S.Sec	C.Sec	Voc	S.Univ	C.Univ	Grad/Prof	PNS
No-B100	0.32%	2.27%	12.99%	11.04%	6.49%	42.53%	24.35%	0.00%
No-B320	0.33%	0.00%	17.67%	13.33%	7.33%	39.33%	21.67%	0.33%
Hyp-B100-P10	0.00%	1.00%	14.00%	16.67%	5.33%	41.33%	21.67%	0.00%
Hyp-B100-P100	0.00%	0.66%	17.61%	14.29%	5.32%	41.86%	20.27%	0.00%
Hyp-B320-P100	0.33%	0.33%	16.07%	11.48%	8.85%	39.02%	23.61%	0.33%
Real-B100-P1	0.00%	0.33%	14.90%	10.60%	10.26%	43.71%	19.54%	0.66%
Real-B100-P10	0.00%	0.00%	18.67%	13.00%	7.00%	36.67%	24.00%	0.67%
Total	0.14%	0.66%	15.97%	12.90%	7.23%	40.64%	22.16%	0.28%

p-value (chi-squared test) = 0.130

Panel D: Household Income

	< 20,000	20,000–39,999	40,000–59,999	60,000–99,999	> 100,000
No-B100	12.34%	29.55%	20.78%	28.25%	9.09%
No-B320	14.33%	33.00%	26.33%	18.33%	8.00%
Hyp-B100-P10	13.33%	31.00%	21.33%	26.33%	8.00%
Hyp-B100-P100	9.97%	29.90%	24.25%	27.57%	8.31%
Hyp-B320-P100	15.41%	37.38%	19.02%	18.69%	9.51%
Real-B100-P1	13.58%	37.42%	21.85%	21.19%	5.96%
Real-B100-P10	11.33%	30.33%	26.33%	25.33%	6.67%
Total	12.90%	32.66%	22.83%	23.68%	7.94%

p-value (chi-squared test) = 0.070

A.2 Pre-registered Treatment Comparisons

Table A3 presents an overview of all pre-registered treatment comparisons, including differences in means, Cohen's d effect sizes, and p-values from two-sided Wilcoxon rank-sum tests.

Table A3: Pre-registered Treatment Comparisons - Decoding Scores

Treatment Comparison	**Diff. in Means**	**Cohen's** d	p**-value**	**Study**
Real–B100–P10 vs. No–B100	13.1	1.66	<0.001	Study 1
Real–B100–P10 vs. Hyp–B100–P10	11.7	1.41	<0.001	Study 1
Hyp–B100–P10 vs. No–B100	1.5	0.16	0.091	Study 1
Real–B100–P10 vs. No–B100	15.6	2.30	<0.001	Study 2
Real–B100–P10 vs. Hyp–B100–P10	14.9	2.09	<0.001	Study 2
Hyp–B100–P10 vs. No–B100	0.7	0.09	0.747	Study 2
No–B320 vs. No–B100	3.3	0.39	<0.001	Study 2
Hyp–B100–P100 vs. Hyp–B100–P10	0.9	0.10	0.256	Study 2
Hyp–B320–P100 vs. No–B320	2.2	0.23	0.005	Study 2
Real–B100–P1 vs. No–B320	4.4	0.44	<0.001	Study 2

Notes: Each row reports a pre-registered treatment contrast of decoding scores between two experimental conditions. All effects that are statistically significant at the 5% level remain so after Bonferroni correction for multiple comparisons. However, the difference between *Hyp–B100–P10* and *No–B100* in Study 1 is no longer significant at the 10% level.

A.3 Alternative Effort Measures

In Section 4, we use the number of correctly solved items as our primary measure of effort, e, because the decoding task requires relatively low skill. Participants who dedicated a reasonable amount of time to an item almost always solved it correctly. As such, this measure closely reflects genuine effort.

In this section, we additionally consider two alternative effort measures: the number of *attempted* items (regardless of correctness) and the *time spent* on the decoding task. While we discuss below why these measures may be less precise than the number of correctly solved items, we also demonstrate that our findings are robust to their use.

The number of attempted items is affected by a design feature of the experiment: participants were allowed to proceed without correctly solving any item. Consequently, individuals who rapidly clicked through the task without exerting meaningful effort could appear to have high effort under this metric. Time spent on the decoding task is also less suitable because, without a time limit, completion times varied substantially, including some notable outliers.

Figures A1 and A2 show the cumulative distribution functions (CDFs) for all three effort measures across both studies. Tables A4–A5 report OLS regression results using the alternative measures for both Study 1 (Table A4) and Study 2 (Table A5), with and without demographic controls. The results are highly consistent across studies and irrespective of how effort is measured.

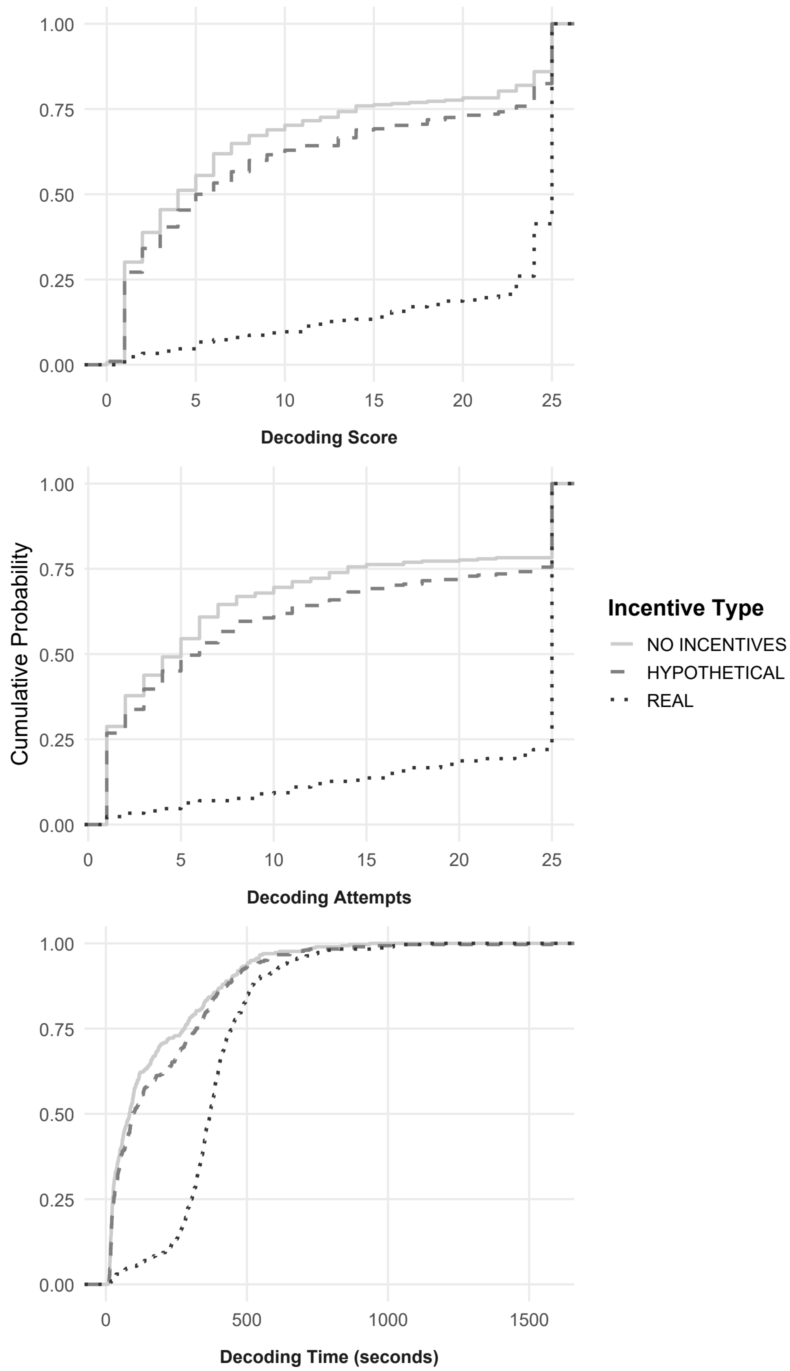


Figure A1: Cumulative distribution functions of effort provision across treatments - Study 1. Line styles indicate incentive type: solid (No-Incentives), dashed (Hypothetical), and dotted (Real).

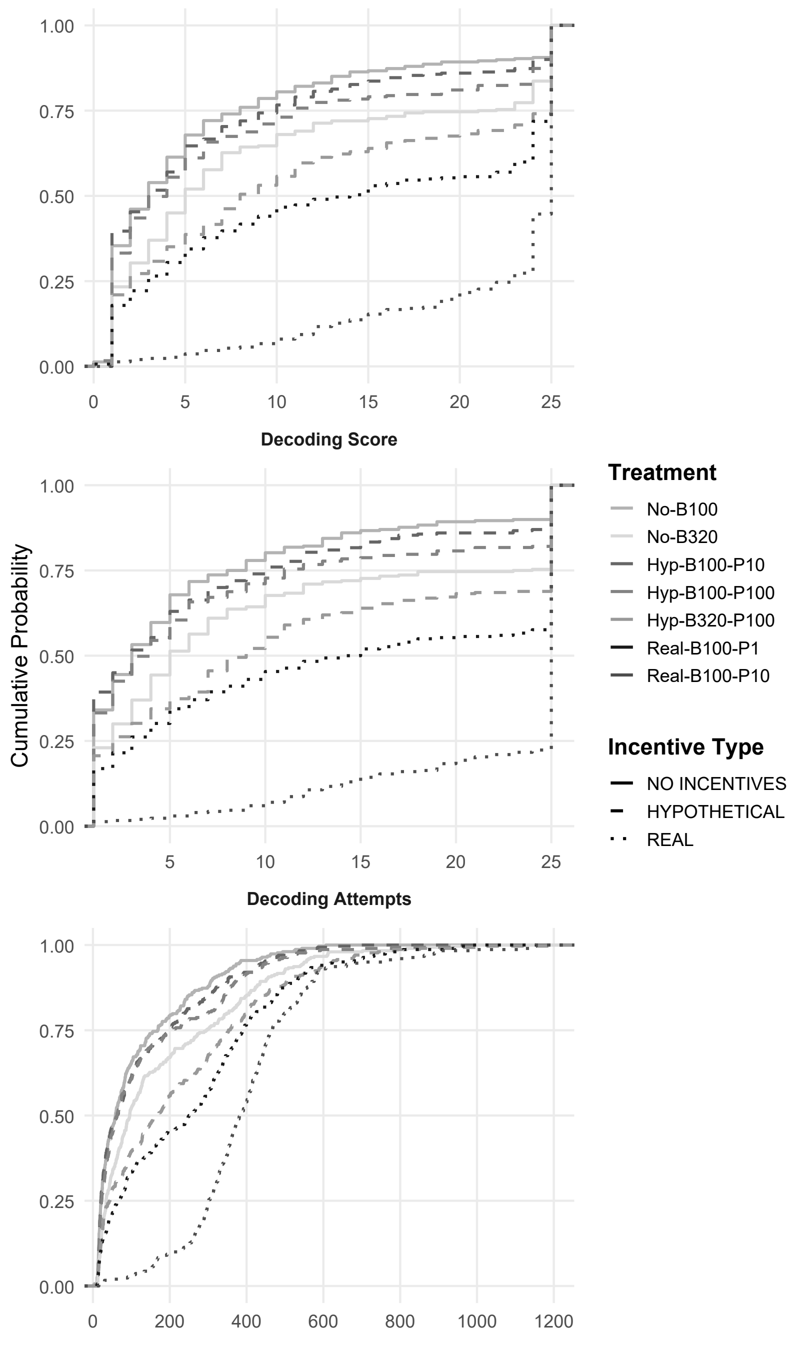


Figure A2: Cumulative distribution functions of effort provision across treatments - Study 2. Line styles indicate incentive type: solid (No-Incentives), dashed (Hypothetical), and dotted (Real). Within each incentive type, grayscale shading reflects relative performance, with darker shades indicating higher mean decoding scores.

Table A4: Regression Analysis of Effort Provision in Study 1 (OLS)

Dependent Variable:	Decoding Scores		Decoding Attempts		Decoding Time (sec.)	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment (ref = No-B100)						
Hyp-B100-P10	1.487*	1.390*	1.389*	1.296*	27.197*	27.951*
	(0.768)	(0.773)	(0.781)	(0.787)	(16.092)	(16.165)
Real-B100-P10	13.145***	13.069***	13.224***	13.156***	216.887***	215.105***
	(0.648)	(0.646)	(0.656)	(0.653)	(14.267)	(14.120)
Age (ref = Under 24)						
25–34 years		1.461		1.609		21.307
		(1.034)		(1.049)		(25.631)
35–44 years		0.457		0.509		-16.262
		(1.056)		(1.073)		(24.919)
45–54 years		1.205		1.241		21.917
		(1.107)		(1.121)		(26.995)
55–64 years		-0.620		-0.648		-10.497
		(1.231)		(1.243)		(27.331)
65+ years		-1.055		-1.029		6.660
		(1.574)		(1.596)		(32.899)
Gender (ref = Male)						
Female		-0.287		-0.342		-11.751
		(0.586)		(0.594)		(12.970)
Non-binary / third		-3.507		-3.731		-93.667
		(4.563)		(4.566)		(75.400)
Prefer to self-describe		-12.007***		-12.225***		-241.046***
		(4.363)		(4.404)		(59.022)
Prefer not to say		0.103		0.300		62.030
		(5.082)		(4.973)		(135.345)
Income (ref = <£20k)						
£20k–39,999		1.450		1.523		16.966
		(0.947)		(0.954)		(21.213)
£40k–59,999		0.513		0.602		1.614
		(1.047)		(1.057)		(23.763)
£60k–99,999		1.346		1.412		-14.779
		(1.038)		(1.042)		(22.320)
£100k+		1.350		1.396		0.305
		(1.380)		(1.382)		(30.200)
Education (ref = Primary)						
Some Secondary		-0.244		0.321		11.800
		(5.816)		(5.887)		(82.402)
Completed Secondary School		0.929		1.092		49.197
		(5.661)		(5.689)		(77.546)
Vocational / Similar		-1.255		-1.062		-17.053
		(5.691)		(5.721)		(77.902)
Some University (no degree)		-1.678		-1.482		-21.896
		(5.724)		(5.755)		(79.761)
Bachelor's Degree		-0.847		-0.704		0.812
		(5.670)		(5.698)		(77.832)
Graduate / professional		-2.503		-2.368		-22.898
		(5.691)		(5.721)		(78.300)
Prefer not to say		-6.020		-6.290		-66.493
		(8.226)		(8.211)		(162.460)
Constant	8.615***	8.153	8.890***	8.169	163.951***	161.393**
	(0.528)	(5.768)	(0.540)	(5.799)	(10.666)	(80.799)
Observations	901	901	901	901	901	901
R-squared	0.321	0.344	0.320	0.343	0.211	0.241

Notes: Robust standard errors in parentheses. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Reference categories are shown in italics.

Table A5: Regression Analysis of Effort Provision in Study 2 (OLS)

Dependent Variable:	Decoding Scores		Decoding Attempts		Decoding Time (sec.)	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment (ref = No-B100)						
No-B320	3.284***	3.226***	3.347***	3.295***	64.770***	61.839***
	(0.689)	(0.690)	(0.698)	(0.699)	(13.229)	(13.083)
Hyp-B100-P10	0.694	0.587	0.701	0.594	14.711	10.136
	(0.635)	(0.640)	(0.641)	(0.646)	(11.244)	(11.368)
Hyp-B100-P100	1.578**	1.501**	1.591**	1.509**	26.739**	24.047**
	(0.666)	(0.661)	(0.673)	(0.669)	(12.250)	(12.091)
Hyp-B320-P100	5.452***	5.402***	5.522***	5.468***	112.105***	109.553***
	(0.704)	(0.707)	(0.708)	(0.711)	(14.195)	(13.982)
Real-B100-P1	7.634***	7.587***	7.803***	7.748***	146.088***	142.252***
	(0.726)	(0.732)	(0.733)	(0.740)	(14.201)	(14.208)
Real-B100-P10	15.567***	15.528***	15.944***	15.897***	287.007***	284.280***
	(0.547)	(0.558)	(0.544)	(0.555)	(12.652)	(12.617)
Age (ref = 18–24)						
25–34 years		1.024		0.932		-8.210
		(0.772)		(0.784)		(17.095)
35–44 years		0.286		0.183		-7.318
		(0.769)		(0.781)		(17.180)
45–54 years		-0.158		-0.287		1.349
		(0.792)		(0.803)		(17.880)
55–64 years		1.043		1.080		45.924**
		(0.828)		(0.836)		(19.022)
65+ years		1.453		1.401		55.053**
		(1.024)		(1.042)		(22.171)
Gender (ref = Male)						
Female		0.040		-0.049		-4.586
		(0.379)		(0.382)		(7.781)
Non-binary / third		0.074		0.049		-1.896
		(2.597)		(2.583)		(42.289)
Prefer to self-describe		-7.171***		-7.394***		-158.013***
		(1.007)		(0.981)		(28.485)
Prefer not to say		-3.182**		-2.761*		52.240
		(1.541)		(1.456)		(118.211)
Income (ref = <£20k)						
£20k–39,999		-1.237*		-1.263*		-13.522
		(0.661)		(0.669)		(13.494)
£40k–59,999		-0.841		-0.900		-23.179
		(0.706)		(0.715)		(14.146)
£60k–99,999		-1.455**		-1.386**		-38.605***
		(0.691)		(0.700)		(13.633)
£100k+		-1.942**		-1.974**		-56.679***
		(0.869)		(0.883)		(16.524)
Education (ref = Primary)						
Some Secondary		7.332**		6.696**		-3.707
		(2.955)		(2.671)		(105.832)
Completed Secondary School		8.044***		7.626***		18.652
		(1.974)		(1.510)		(98.133)
Vocational / Similar		8.261***		7.683***		27.748
		(1.995)		(1.537)		(98.283)
Some University (no degree)		6.120***		5.692***		-26.323
		(2.032)		(1.589)		(98.356)
Bachelor's Degree		7.547***		7.038***		12.835
		(1.939)		(1.463)		(97.861)
Graduate / professional		8.170***		7.725***		26.093
		(1.964)		(1.496)		(98.150)
Prefer not to say		8.604**		7.841**		11.622
		(3.728)		(3.483)		(126.570)
Constant	6.266***	-0.898	6.403***	-0.151	114.170***	119.712
	(0.428)	(2.118)	(0.430)	(1.698)	(7.422)	(98.737)
Observations	2,116	2,116	2,116	2,116	2,116	2,116
R-squared	0.253	0.263	0.259	0.269	0.214	0.239

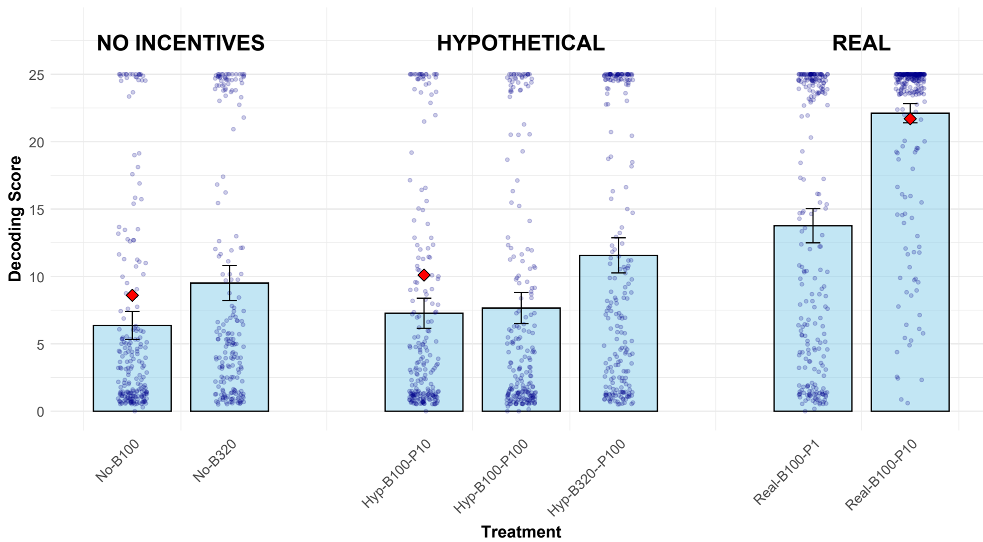
Notes: Robust standard errors in parentheses. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Reference categories are shown in italics.

A.4 Error Analysis

A.4.1 Comprehension Errors and Effort Provision

The main objective of the paper is to identify how effort responds to incentive structures. Therefore, participants needed to clearly understand the incentive scheme. Figure A3 replicates Figure 2 for the subset of participants who made no comprehension errors, and it looks nearly identical. This provides initial evidence that mistakes in the comprehension question do not drive our treatment effects.

Figure A3: Decoding Scores by Treatment for Participants Without Errors



We incorporated two key design features to ensure that participants understand the incentive scheme. First, participants were required to correctly answer the comprehension question, ensuring that any initial misunderstanding was corrected before proceeding to the real-effort task. Second, after each submitted answer during the task, participants were reminded of their incentive scheme before deciding whether to continue, reinforcing understanding throughout the experiment.

The regression analysis in Table A6 further confirms the effectiveness of these measures. Column (1) regresses effort provision on treatment indicators and replicates the treatment effects reported in the main analysis. Column (2) includes a dummy for participants who initially failed the comprehension question and Column (3) fully interacts it with the treatment indicators. The comprehension-error dummy is statistically insignificant in Column (2) and Column (3), and none of the interaction terms are significant at the 5% level in Column (3). Most importantly, the treatment coefficients remain highly stable across specifications and relative to one another when comprehension errors are included in Column (2) and Column (3).

Table A6: Regression of Decoding Scores on Treatments and Error Interactions

Dependent Variable:	**Decoding Scores**				
	(1)	(2)	(3)	(4)	(5)
No-B320	3.284***	3.287***	3.155***	2.962***	3.421***
	(0.689)	(0.689)	(0.852)	(0.828)	(0.731)
Hyp-B100-P10	0.694	0.692	0.918	0.640	0.900
	(0.635)	(0.635)	(0.778)	(0.760)	(0.671)
Hyp-B100-P100	1.578**	1.575**	1.300	1.144	1.738**
	(0.666)	(0.667)	(0.795)	(0.776)	(0.704)
Hyp-B320-P100	5.452***	5.449***	5.206***	4.720***	5.705***
	(0.704)	(0.705)	(0.849)	(0.822)	(0.750)
Real-B100-P1	7.634***	7.624***	7.404***	7.321***	7.621***
	(0.726)	(0.729)	(0.837)	(0.819)	(0.770)
Real-B100-P10	15.567***	15.554***	15.757***	15.530***	15.768***
	(0.547)	(0.553)	(0.643)	(0.633)	(0.575)
Comprehension Error		-0.093	-0.283		
		(0.432)	(0.898)		
Incentive Error				-0.833	
				(0.893)	
Task Error					1.200
					(1.567)
No-B320 × Error			0.396	1.105	-1.427
			(1.450)	(1.493)	(2.250)
Hyp-B100-P10 × Error			-0.786	0.019	-1.987
			(1.342)	(1.371)	(2.129)
Hyp-B100-P100 × Error			0.941	1.675	-1.654
			(1.468)	(1.550)	(2.214)
Hyp-B320-P100 × Error			0.838	3.068*	-2.424
			(1.527)	(1.613)	(2.248)
Real-B100-P1 × Error			0.983	1.371	-0.295
			(1.765)	(1.977)	(2.367)
Real-B100-P10 × Error			-1.359	-1.273	-2.007
			(1.326)	(1.543)	(1.904)
Constant	6.266***	6.296***	6.355***	6.507***	6.157***
	(0.428)	(0.452)	(0.529)	(0.525)	(0.446)
Observations	2,116	2,116	2,116	2,116	2,116
R-squared	0.253	0.253	0.254	0.255	0.254

Notes: Robust standard errors in parentheses. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A more granular analysis of initial failures on the comprehension check further reinforces this conclusion. We separate the comprehension question into three incentive-related items (i.e., those concerning the incentive scheme and participation fee) and three task-related items (i.e., those concerning the maximum number of decoding tasks and the continuation rule) and generate dummy variables for mistakes in these two error categories. Columns (4) and (5) repeat the analysis in

Column (3), distinguishing between errors on incentive-related and task-related items. Again, treatment coefficients are largely unaffected, and none of the interaction terms reach significance at the 5% level.

Overall, these results show that initial misunderstanding did not influence effort provision and that the treatment effects in our effort measure are not driven by differential comprehension across conditions.

A.4.2 Treatments and Comprehension Errors

Section 4.1 shows at the treatment-category level (No-Incentives, Hypothetical, and Real) that non-incentivized conditions exhibit a substantially higher share of participants who failed the comprehension question on their first attempt. In the analysis that follows, we further unpack these results by disaggregating them across the seven treatments and differentiating between types of errors. Specifically, we categorize the answer options in the comprehension question into three incentive-related items (i.e., those concerning the incentive scheme and participation fee) and three task-related items (i.e., those concerning the maximum number of decoding tasks and the continuation rule). Figure A4 reports these results.

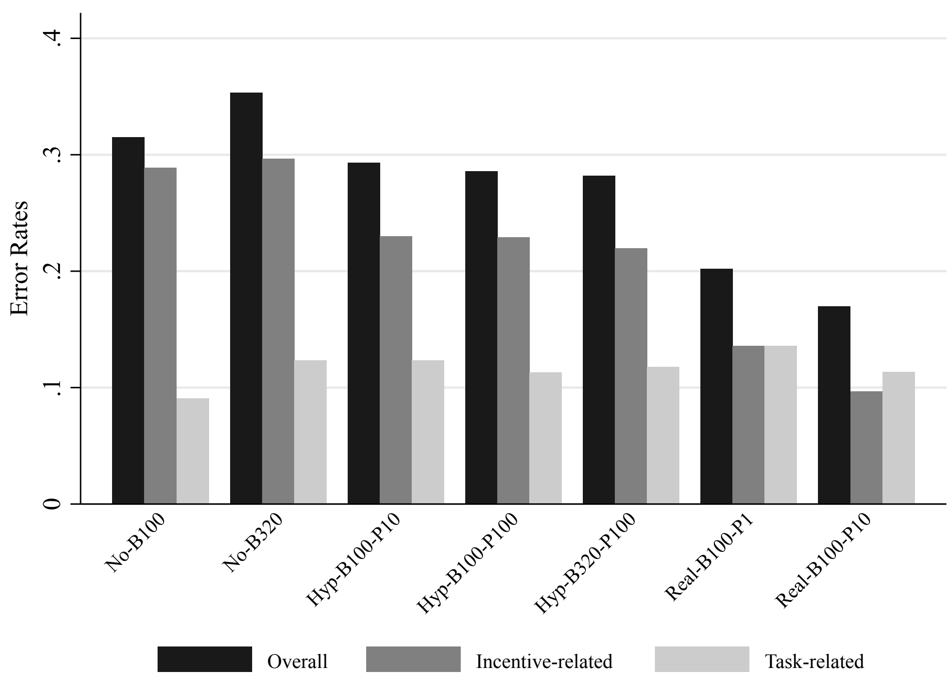


Figure A4: Error Rates on First Attempts

Consistent with the findings in Section 4.1, overall error rates differ significantly across treatments ($p < 0.001$, χ^2-test), with the lowest error rate observed in the treatment Real-B100-P10.[8] These treatment differences are driven by errors related to the incentive-related items rather than the task-related items. We observe statistically significant variation in incentive-related errors ($p < 0.001$, χ^2-test), whereas task-related errors show no significant differences ($p = 0.759$, χ^2-test).

When clustering treatments into categories, failure rates in incentive-related questions differ substantially across treatment categories: 29.3% in No-Incentives, 22.7% in Hypothetical, and 11.6% in Real (χ^2-test: $p < 0.001$ for Real vs. both other categories; Hypothetical vs. No-Incentives, $p = 0.004$).[9] In contrast, no statistically significant differences are observed for task-related errors (χ^2-test: $p > 0.33$ for all pairwise category comparisons).

These findings raise the question of whether treatment differences in comprehension errors reflect inherent difficulty in understanding certain incentive structures or instead a lack of engagement resulting from the absence of incentives. While the significant differences in incentive-related errors and the non-significant differences in task-related errors could be superficially interpreted as evidence that confusion about incentives drives our error patterns, we argue that a lack of engagement is at least equally well supported by our data and features of our design:

1. Difficulty of the comprehension question: If differences in incentive-related errors were primarily due to confusion or difficulty understanding the underlying incentive structures, we would expect a different ordering of No-Incentives and Hypothetical in terms of error rates. No-Incentives features the simplest incentive structure of a fixed participation fee, yet we observe a significantly higher error rate in this condition than in the other two conditions (both comparisons with $p < 0.005$, χ^2-test).

This becomes even more apparent when we take a closer look at the decision screen subjects faced. On the same screen, the instructions and the comprehension question based on them are presented simultaneously. Subjects are first informed:

[8] Within each treatment category, overall, incentive-related, and task-related error rates do not differ significantly (No-Incentives: $p = 0.316$ / $p = 0.835$ / $p = 0.196$; Hypothetical: $p = 0.952$ / $p = 0.944$ / $p = 0.925$; Real: $p = 0.313$ / $p = 0.135$ / $p = 0.405$, all χ^2-tests).

[9] We focus on pairwise comparisons between categories rather than between individual treatments to mitigate concerns about multiple testing.

[. . .] You will receive no bonus payment for solving the decoding tasks.

Immediately afterwards, on the same screen, the very first comprehension question asks them to assess whether the following statement is true:

You will receive no bonus payment for solving the decoding tasks.[10]

Subjects in No-Incentives performed significantly worse on this question, which offered the easiest possible answer option, than subjects in Hypothetical and Real (both comparisons with $p < 0.01$, χ^2-test). It is difficult to argue that these differences arise solely from the difficulty of the question.

2. Task related questions: We would argue that the task related questions were in fact overall simpler and had a lower threshold for engagement. Recall, that on the same screen, the instructions and the comprehension question are presented simultaneously. The first lines of the instructions read:

> *The maximum number of decoding tasks you can solve is 25.*
>
> *After completing the first task, you can decide how much you want to work. After completing each decoding task, you can decide whether you want to continue or stop working.*

Thus, a subject who diligently paid attention to the first three sentences of the instructions obtained all necessary information to correctly answer the task-related questions. Moreover, this information was provided before the instructions began to differ across treatments with respect to the incentive-related information. These task-related items in the comprehension question should therefore be least affected by the treatments, which is consistent with our interpretation that the treatments influenced subjects' engagement with the experiment and the information provided, rather than the difficulty of the questions.

[10] Corresponding statements were in Hypothetical: *You will receive a bonus payment of X pence for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account* and in Real: *You will receive a bonus payment of X pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account.*

A.5 Tobit Regressions

Table A7: Tobit Regressions

	Study 1		Study 2	
	(1)	(2)	(3)	(4)
Treatment (ref = No-B100)				
No-B320			3.759***	3.672***
			(0.876)	(0.877)
Hyp-B100-P10	1.700*	1.535	0.794	0.649
	(0.933)	(0.932)	(0.873)	(0.870)
Hyp-B100-P100			1.731**	1.614*
			(0.874)	(0.871)
Hyp-B320-P100			6.502***	6.426***
			(0.878)	(0.877)
Real-B100-P1			8.926***	8.865***
			(0.883)	(0.885)
Real-B100-P10	18.350***	18.216***	20.376***	20.319***
	(1.014)	(1.003)	(0.936)	(0.938)
25–34 years		1.281		1.440
		(1.531)		(0.993)
35–44 years		0.251		0.526
		(1.543)		(0.998)
45–54 years		1.261		-0.087
		(1.601)		(1.029)
55–64 years		-0.915		1.672
		(1.763)		(1.061)
65+ years		-1.713		1.792
		(2.071)		(1.284)
Female		-0.447		0.137
		(0.799)		(0.480)
Non-binary / third		-4.499		0.527
		(5.129)		(2.847)
Self-describe (PSD)		-14.677*		-8.259
		(7.956)		(6.155)
Prefer not to say		0.723		-2.206
		(5.153)		(4.744)
£20k–39,999		1.923		-1.452*
		(1.329)		(0.791)
£40k–59,999		0.899		-0.888
		(1.404)		(0.847)
£60k–99,999		2.176		-2.018**
		(1.435)		(0.847)
£100k+		1.892		-2.292**
		(1.886)		(1.100)
Some Secondary		0.780		10.806
		(7.365)		(7.109)
Completed Secondary School		3.616		11.506*
		(6.506)		(6.525)
Vocational / Similar		0.841		12.168*
		(6.541)		(6.535)
Some University (no degree)		-0.031		9.124
		(6.621)		(6.560)
Bachelor's Degree		1.269		11.100*
		(6.511)		(6.512)
Graduate / professional		-0.764		11.884*
		(6.556)		(6.521)
Prefer not to say		-6.457		11.916
		(8.988)		(7.972)
Constant	9.294***	6.536	6.591***	-4.239
	(0.661)	(6.624)	(0.613)	(6.589)
Observations	901	901	2,116	2,116
Pseudo R^2	0.063	0.068	0.044	0.047

Notes: The table reports Tobit regression estimates with lower and upper censoring at 0 and 25, respectively. Standard errors in parentheses. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.6 Budget-Constrained Counterfactuals

This appendix derives the counterfactual effort curves shown in Figure A5, plotting predicted effort against total per-participant cost under real piece-rate incentives, hypothetical piece-rate incentives, and base pay only.

Let B denote base pay in pounds. The paper already establishes that non-pecuniary motivation s depends on B: our experimental design identifies s at exactly two base-pay levels, $\hat{s}_L$ at $B = £1.00$ and $\hat{s}_H$ at $B = £3.20$ (see Section 3.2). Similarly, our design identifies β at three discrete (B, p) combinations: $\hat{\beta}_{LL}$, $\hat{\beta}_{LH}$, and $\hat{\beta}_{HH}$ (Table 2).

We denote the experimenter's total expected cost per participant by C. Base pay B is always paid; real piece-rate payments cost p per item solved; hypothetical piece rates are costless. Thus:

$$C = B + p \cdot e^{*} \cdot I_{\text{Real}}.$$

Linearity assumptions. Figure A5 plots effort as a continuous function of B, so evaluating s and β between these identified points requires a functional form connecting them. For B within the observed range, $£1.00 \leq B \leq £3.20$, we make two linearity assumptions.

1. *Linear $s(B)$.* We assume non-pecuniary motivation varies linearly between its two identified values. This gives:
$$\hat{s}(B) = \hat{s}_L + \frac{\hat{s}_H - \hat{s}_L}{3.20 - 1.00}(B - 1.00).$$

2. *Linear $\beta(B)$ at $p = £1.00$.* Holding the piece rate fixed at $p = £1.00$, we assume the internalization rate varies linearly in base pay B over the full range $£1.00 \leq B \leq £3.20$, using $\hat{\beta}_{LH}$ (at $B = £1.00$) and $\hat{\beta}_{HH}$ (at $B = £3.20$) as the two anchor points. This gives:
$$\hat{\beta}(B) = \hat{\beta}_{LH} + \frac{\hat{\beta}_{HH} - \hat{\beta}_{LH}}{3.20 - 1.00}(B - 1.00).$$

These two assumptions are imposed solely to construct Figure A5; no other result in the paper relies on them, and Section 3.2 does not assume linearity in B anywhere in the main estimation.

General effort expression. Extending the first-order condition in equation (4) to allow s and β

to vary with B:

$$\hat{k}\, e^{\hat{\gamma}} = \hat{s}(B) + \hat{\beta}(B) \cdot p \cdot I_{\text{Hyp}} + p \cdot I_{\text{Real}},$$

and solving for e^* exactly as equation (3) is derived from equation (2), we obtain:

$$e^* = \left(\frac{\hat{s}(B) + \hat{\beta}(B) \cdot p \cdot I_{\text{Hyp}} + p \cdot I_{\text{Real}}}{\hat{k}} \right)^{1/\hat{\gamma}}.$$

The three counterfactual curves follow by substituting the relevant budget allocation into the budget constraint and effort expression above.

Real piece-rate curve ($I_{\text{Real}} = 1$, $I_{\text{Hyp}} = 0$). We fix $B = £1.00$, so $\hat{s}(B) = \hat{s}_L$ is point-identified and no interpolation is required. Substituting into the effort expression:

$$e^*_{\text{Real}} = \left(\frac{\hat{s}_L + p}{\hat{k}} \right)^{1/\hat{\gamma}}.$$

Substituting into the budget constraint:

$$C = £1.00 + p \cdot \left(\frac{\hat{s}_L + p}{\hat{k}} \right)^{1/\hat{\gamma}}.$$

Define $f(p) \equiv p \cdot \big((\hat{s}_L + p)/\hat{k}\big)^{1/\hat{\gamma}}$. Since f is strictly increasing in p, there exists for every $C > £1.00$ a unique solution $p^*(C)$ to $f(p) = C - £1.00$, which we solve numerically. Predicted effort on the real curve is:

$$e^*_{\text{Real}}(C) = \left(\frac{\hat{s}_L + p^*(C)}{\hat{k}} \right)^{1/\hat{\gamma}}.$$

This curve is built entirely from point-identified parameters ($\hat{k}$, $\hat{\gamma}$, $\hat{s}_L$) and does not rely on the linearity assumptions above.

Hypothetical piece-rate curve ($I_{\text{Hyp}} = 1$, $I_{\text{Real}} = 0$). Since hypothetical piece rates are costless, the budget constraint gives $C = B$, so the entire budget is available for base pay. We fix $p = £1.00$—the highest piece rate tested, providing an upper bound on achievable hypothetical effort at any given budget.[11] Substituting $B = C$ and $p = £1.00$ into the effort expression:

$$e^*_{\text{Hyp}}(C) = \left(\frac{\hat{s}(C) + \hat{\beta}(C) \cdot £1.00}{\hat{k}} \right)^{1/\hat{\gamma}}.$$

Base-pay-only curve ($I_{\text{Real}} = 0$, $I_{\text{Hyp}} = 0$, $p = 0$). Again $C = B$, so:

$$e^*_{\text{Base}}(C) = \left(\frac{\hat{s}(C)}{\hat{k}} \right)^{1/\hat{\gamma}}.$$

Since $\hat{\beta}(C) > 0$ throughout our budget range, $e^*_{\text{Hyp}}(C) > e^*_{\text{Base}}(C)$ at every C, as expected.

[11] This choice is dictated by identification: $p = £1.00$ is the only piece-rate level at which β is observed at both base-pay levels, providing the two anchors $\hat{\beta}_{LH}$ and $\hat{\beta}_{HH}$ required for linear interpolation in B. At $p = £0.10$, β is observed only at $B = £1.00$, leaving the interpolation under-identified: a single anchor point does not determine a line.

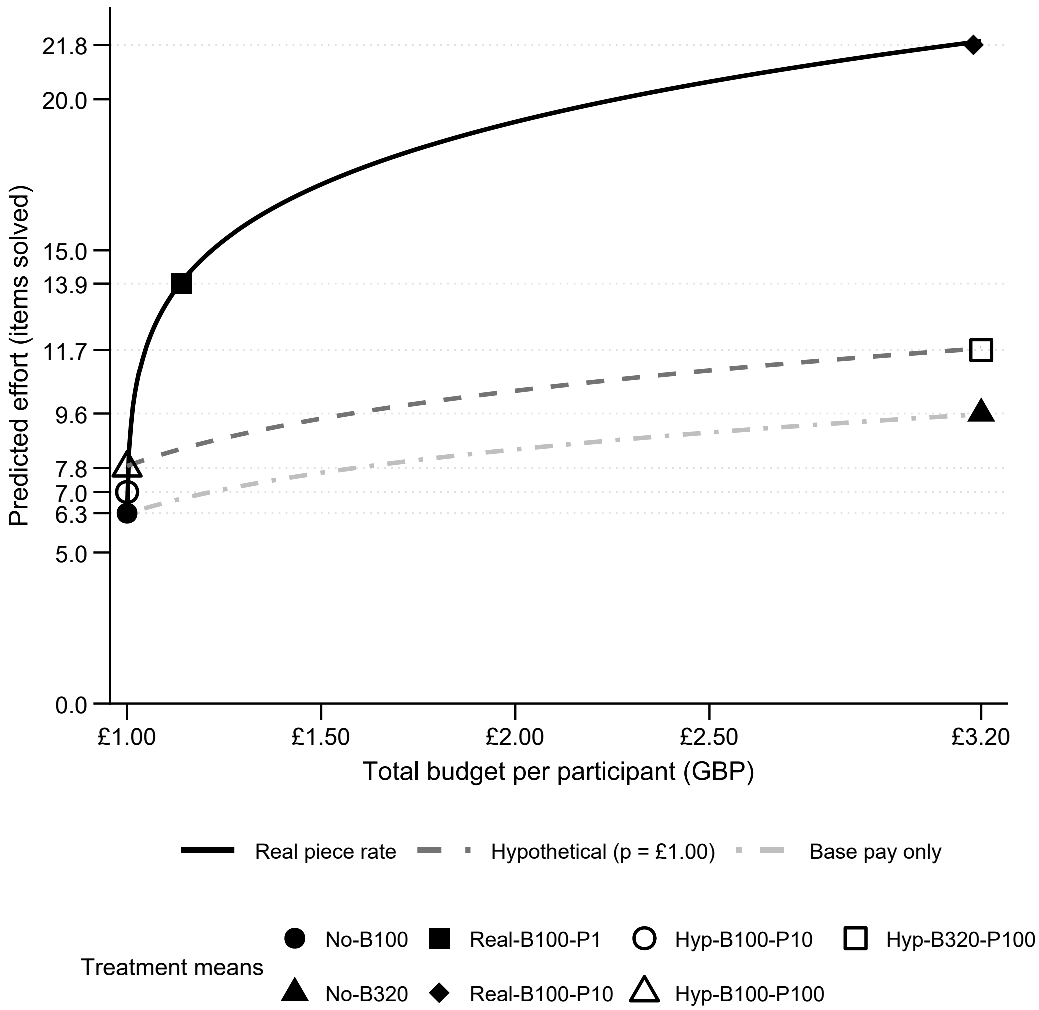


Figure A5: Budget-constrained counterfactual. The x-axis shows total per-participant budget $C \in [\pounds 1.00,\ \pounds 3.20]$; the y-axis shows predicted effort e^*. Treatment means are shown as markers (see legend); those at a curve's anchor points lie on it by construction, not model fit. Hyp-B100-P10 (open circle) falls below the hypothetical curve because it uses $p = \pounds 0.10$, not $\pounds 1.00$.

B Experimental Instructions

Please select "Strongly agree" to show that you are paying attention to this question.

- Strongly agree
- Agree
- Disagree
- Strongly disagree

Welcome to our study! As a reminder, you must complete this survey on a desktop computer. No mobile devices are allowed. Please click next to continue.

Consent Form

Principal Investigator: Sebastian J. Goerg (Technical University of Munich)
Co-investigator: Orestis Kopsacheilis (Technical University of Munich)
Co-investigator: Christoph Drobner (Technical University of Munich)

You are invited to participate in a research study about decision-making. If you agree to be part of the research study, you will be asked to make decisions on your computer.

Data collected during this study will be linked only to your Prolific ID, and in no way will it be linked to your name. Therefore, your behavior during this study is completely anonymous.

Participating in this study is completely voluntary. If you have questions about this research study, please contact Christoph Drobner (christoph.drobner@tum.de).

The study is a no-deception study. All the instructions and information you will receive in this study are true and accurate.

Please click the arrow button if you attest to being at least 18 years old and agree to take part in the study. Please close the browser window if you are not willing to take part in the study.

In this study, you can work on a series of decoding tasks.

One such decoding task is illustrated below. Your task is to decode text using a number key that is provided in the bottom of the screen and enter the answer into an input field. In this example, the correct answer is WKBLE (the non-capitalized version "wkble" also counts as correct). This solution is achieved by looking up the corresponding letter for each number in the panel.

W	U	H	J	P	L	E	B	Z	K
9	5	2	4	6	3	7	8	0	1

91837

To gain experience with this task, we ask you to solve one example on the following page. In this example, you will receive feedback whether you solved the task correctly but during the main task you will receive no feedback.

Please decode the number into text. This is achieved by looking up the corresponding letter for each number.

Q	U	O	W	F	C	M	I	Z	J
8	9	6	0	4	2	5	7	1	3

12584

IF PARTICIPANT FAILED TO CORRECTLY SOLVE THE PRACTICE DECODING TASK:

You did not solve the decoding task correctly. Please try again by decoding the number into text. This is achieved by looking up the corresponding letter for each number.

Q	U	O	W	F	C	M	I	Z	J
8	9	6	0	4	2	5	7	1	3

12584

IF PARTICIPANT CORRECTLY SOLVED THE PRACTICE DECODING TASK:

Correct! You can now proceed with the study.

IF PARTICIPANT WAS ASSIGNED TO ONE OF THE HIGH BASE PAY TREATMENTS:

On top of the advertised £1 you are guaranteed to receive an additional £2.2 for participating in this study. Therefore, irrespective of the choices and answers you give during this study, you are guaranteed to receive a total of £3.2 as long as you submit your completion code.

IF PARTICIPANT WAS ASSIGNED TO ONE OF THE LOW BASE PAY TREATMENTS:

You receive £1 for participating in this study. Therefore, irrespective of the choices and answers you give during this study, you are guaranteed to receive a total of £1 as long as you submit your completion code.

INSTRUCTIONS IN "RED" VARY BETWEEN TREATMENTS

Please read the following instructions carefully.

The maximum number of decoding tasks you can solve is 25.

After completing the first task, you can decide how much you want to work. After completing each decoding task, you can decide whether you want to continue or stop working. You will receive no feedback whether you solved the task correctly. You will receive the participation fee of £[1, 3.2] regardless of when you decide to stop working. You will receive [no bonus payment for solving the decoding tasks; a bonus payment of 10 pence for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account; a bonus payment of £1 for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account; a bonus payment of 10 pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account; a bonus payment of 1 pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account].

Please answer the following comprehension question. Which of the following statements are true? (please select all correct answers)

- You will receive [no bonus payment for solving the decoding tasks; a bonus payment of 10 pence for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account; a bonus payment of £1 for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account; a bonus payment of 10 pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account; a bonus payment of 1 pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account].
- After completing each decoding task, you can decide whether you want to continue or stop working.
- You will only receive the participation fee of £[1, 3.2] if you solve all decoding tasks.
- You will receive the participation fee of £[1, 3.2] regardless of when you decide to stop working.
- The maximum number of decoding tasks that you can solve is 10.
- The maximum number of decoding tasks that you can solve is 25.

IF PARTICIPANT FAILED TO CORRECTLY SOLVE THE COMPREHENSION CHECK QUESTION ABOUT TASK AND INCENTIVES:

You either selected an incorrect answer or did not select a correct one. Please carefully review the instructions below and answer the comprehension question again.

Please answer the following comprehension question. Which of the following statements are true? (please select all correct answers)

- You will receive [no bonus payment for solving the decoding tasks; a bonus payment of 10 pence for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account; a bonus payment of £1 for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account; a bonus payment of 10 pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account; a bonus payment of 1 pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account].
- After completing each decoding task, you can decide whether you want to continue or stop working.
- You will only receive the participation fee of £[1, 3.2] if you solve all decoding tasks.
- You will receive the participation fee of £[1, 3.2] regardless of when you decide to stop working.
- The maximum number of decoding tasks that you can solve is 10.
- The maximum number of decoding tasks that you can solve is 25.

IF PARTICIPANT CORRECTLY SOLVED THE COMPREHENSION CHECK QUESTION ABOUT TASK AND INCENTIVES:

Thank you. You have answered the comprehension question correctly. Please click the arrow button when you are ready to start working.

Please decode the number into text. This is achieved by looking up the corresponding letter for each number.

S	X	A	J	Z	L	N	G	M	F
8	3	5	2	9	0	4	6	7	1
38921									

Recall that you will receive [no bonus payment for solving the decoding tasks; a bonus payment of 10 pence for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account; a bonus payment of £1 for each correctly solved decoding task. This bonus payment is only hypothetical and will not be credited to your Prolific account; a bonus payment of 10 pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account; a bonus payment of 1 pence for each correctly solved decoding task. This bonus payment will be credited to your Prolific account].

Do you want to continue or stop?

- Continue working
- Stop working

Thank you. Please take your time to answer the following questions.

How old are you?
- Under 18
- 18-24 years old
- 25-34 years old
- 35-44 years old
- 45-54 years old
- 55-64 years old
- 65+ years old

How do you describe yourself?
- Male
- Female
- Non-binary / third gender
- Prefer to self-describe:
- Prefer not to say

What is the highest level of education you have completed?
- Some Primary
- Completed Primary School
- Some Secondary
- Completed Secondary School
- Vocational or Similar
- Some University but no degree
- University Bachelor's Degree
- Graduate or professional degree (MA, MS, MBA, PhD, JD, MD, DDS)
- Prefer not to say

What was your total household income before taxes during the past 12 months?
- Less than 20,000 pounds
- 20,000-39,999 pounds
- 40,000-59,999 pounds
- 60,000-99,999 pounds
- More than 100,000 pounds

Please enter your Prolific ID: