The Algorithmic Assignment of Incentive Schemes*

Saskia Opitz[†] Dirk Sliwka[‡] Timo Vogelsang[§] Tom Zimmermann[¶]

This Version: September 6, 2023

Abstract

The assignment of individuals with different observable characteristics to different treatments is a central question in designing optimal policies. We study this question in the context of increasing workers' performance via targeted incentives, using machine learning algorithms with worker demographics, personality traits, and preferences as input. Running two large-scale experiments we show that (i) performance can be predicted by accurately measured worker characteristics, (ii) a machine learning algorithm can detect heterogeneity in responses to different schemes, (iii) a targeted assignment of schemes to individuals increases performance significantly above the level of the single best scheme, and (iv) algorithmic assignment is more effective for workers who have a high likelihood to repeatedly interact with the employer, or who provide more consistent survey answers.

Keywords: Randomized Controlled Trial, Incentives, Heterogeneity, Treatment Effects, Selection, Algorithm, Machine Learning

JEL classification: C21, C93, M52

*We thank Stefano DellaVigna, Jonathan de Quidt, and Matthias Heinz for their feedback and comments. Moreover, we thank participants of the Young ECONtribute Program seminar, the Annual Meeting of the Committee for Organizational Economics 2022, and the Annual SIOE Conference 2022 for helpful comments. We further thank Devin G. Pope for the provision of the code for the real-effort-task, Fabian Meeßen for excellent research assistance, and Richard Guse for technical support. The project was approved by an IRB board. The experiment is registered with the IDs AEARCTR-0008212 and AEARCTR-0008440. The project received funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy – EXC 2126/1– 390838866.

[†]University of Cologne. Faculty of Management, Economics and Social Sciences. Department of Corporate Development. Email: opitz@wiso.uni-koeln.de

[‡]University of Cologne. Faculty of Management, Economics and Social Sciences. Department of Corporate Development. Email: sliwka@wiso.uni-koeln.de

[§]Frankfurt School of Finance & Management. Department of Accounting. Email: t.vogelsang@fs.de

[¶]University of Cologne. Faculty of Management, Economics and Social Sciences. Email: tom.zimmermann@uni-koeln.de

1 Introduction

To motivate employees, employers can choose from a range of different incentive schemes.[1] However, while one person may perform best under, for example, a performance pay scheme, others may be motivated more effectively through different types of incentives. This paper investigates the extent to which worker performance can be improved by targeted assignment of incentive schemes. Concretely, we propose an algorithm that estimates optimal assignment based on individual worker characteristics as inputs.

Recently advanced methods that combine machine learning and modern causal inference hold promise to identify relevant drivers for targeting policies that can be used to improve desired policy outcomes (Wager and Athey 2018; Chernozhukov et al. 2018; Hitsch and Misra 2018; Farrell et al. 2021a; Farrell et al. 2021b). The underlying idea in this growing literature is to move beyond the identification of average treatment effects and move towards identifying conditional average treatment effects for specific individuals. However, although these methods have been applied in illustrative examples in observational data, their merit in using them to determine optimal treatment assignment in various contexts is still a largely unanswered question.

To study the potential of targeted incentive schemes for performance improvements, we ran two consecutive large-scale real-effort experiments with around 12,000 workers on Amazon MTurk.[2] In both experiments, we hired workers for a real-effort task developed by DellaVigna and Pope (2018). The key questions are whether (i) a machine learning algorithm trained with data on individual characteristics can detect heterogeneous responses to different types of incentives, and (ii) to what extent a targeted assignment of incentive schemes by this algorithm in a second experiment can raise performance.

The project proceeds as follows: In the first step, we conducted an initial experiment (Experiment 1) to test the effectiveness of six different incentive schemes for the same real-effort task. The schemes were mainly based on a previous large-scale study by DellaVigna and Pope (2018) and included a fixed wage, a piece rate scheme, two target bonus schemes with either a gain or loss framing, a competitive scheme with real-time rank feedback, and a social incentive scheme combining a piece rate with a performance-contingent donation to charity. Prior to assigning participants to a scheme, we elicited detailed survey information on subjects' characteristics such as their demographics, personality traits, and their social and economic preferences.

[1] Many studies have shown positive average performance effects of different specific incentive schemes. This includes studies on performance pay (e.g., Lazear 2000; Bandiera et al. 2007; Manthei et al. 2023), tournaments (e.g., Casas-Arce and Martinez-Jerez 2009; Delfgaauw et al. 2013), team incentives (e.g., Friebel et al. 2017), gain-framed incentives and loss-framed incentives (e.g., Hossain and List 2012; Levitt et al. 2016), relative performance evaluation (e.g., Blanes i Vidal and Nossol 2011; Eyring and Narayanan 2018; Barankay 2012), or social incentives (e.g., Imas 2014; Tonin and Vlassopoulos 2015; Gosnell et al. 2020). For a more complete overview see, e.g., Bandiera et al. (2011), Sprinkle and Williamson (2006), Lazear (2018).

[2] The project was approved by an IRB board. The experiment is registered with the IDs AEARCTR-0008212 and AEARCTR-0008440.

Comparing point estimates of the treatment effects, the highest average performance in Experiment 1 is achieved by a scheme that awards a bonus that is lost when the worker fails to achieve a specific target value (*Bonus Loss*).[3] However, when estimating conditional average treatment effects, we detected significant heterogeneity based on worker characteristics in the data. In other words, our estimated model predicted that for subsets of workers of different characteristics, different schemes would lead to higher performance.

We validated this prediction in a second experiment (Experiment 2) where we again elicited the respective workers' characteristics. In Experiment 2, we compare three treatments: (i) a control treatment where all workers received a fixed wage, (ii) a *Best ATE* treatment where all workers received the scheme that generated the highest average treatment effect in Experiment 1 (*Bonus Loss*), and (iii) an *Algorithm* treatment where workers were assigned to the scheme that was predicted to yield the highest performance conditional on their specific characteristics.

In addition to standard tuning of algorithm-specific hyperparameters, we determined the optimal subset of incentive schemes to be implemented in Experiment 2 by maximizing the predicted treatment effect. The resulting set of incentive schemes from this procedure includes the benchmark *Bonus Loss* scheme, a competitive *Real-time Rank Feedback* condition where subjects' pay is based on their prospective percentage rank, and the *Social PfP* scheme where subjects receive a piece rate topped up by a performance-contingent donation to a charity.

We found that the treatment with the targeted scheme assignment significantly outperforms the loss treatment, which had achieved the highest treatment effect in Experiment 1. Specifically, while the loss treatment raised performance by 23.9% over the level of the fixed wage control group, the targeted assignment raised performance by 29.3%, indicating a 5.4 percentage point or 22.5% higher treatment effect.

We furthermore find that the average treatment effect is primarily driven by participants who had previously taken part in Experiment 1 ("Retakers"). Therefore, it is crucial to investigate why the algorithmic assignment was less effective for the subjects who did not participate in Experiment 1 ("New Hires"). We consider four different potential explanations, and provide evidence that neither the prior experience with the task nor differences in observable covariates between New Hires and the algorithm's training sample (referred to as covariate shift in the machine learning literature) can account for the difference in the effectiveness of the assignment algorithm in the two sub-samples. But we find that New Hires who are consistent in their survey responses and who are likely to engage with the platform in the future (as assessed by a predictive ML algorithm), display a treatment effect of similar magnitude to the treatment effect among Retakers. This indicates that the discrepancy in treatment effects can be attributed to inaccurate measurement of traits among "one-shot" participants and other unobservable differences between subjects who are likely to repeatedly offer their services on the platform and those one-shot participants.

[3]This scheme is not the highest performing scheme in DellaVigna and Pope (2018) where the high incentive gain scheme has a higher average treatment effect, but the difference is small (around 1.5%) and not statistically significant. Given the same monetary incentive size, the loss-framed incentive has a non-significantly higher point estimate than the gain-framed incentive. Several studies find a larger performance effect for loss-framed incentives compared to gain-framed incentives (e.g., Hannan et al. 2005; Armantier and Boly 2015; Imas et al. 2017; Van der Stede et al. 2020; Fryer et al. 2022). Others do not find a statistically significant difference (e.g., Grolleau et al. 2016; Levitt et al. 2016; De Quidt et al. 2017; Czibor et al. 2022) or mixed results (e.g., Hossain and List 2012). See Ferraro and Tracy (2022) for a meta-analysis.

Our study contributes to various strands of the literature. We contribute to the literature on the heterogeneous effects of incentive schemes. Previous studies have found heterogeneity in the effect of incentive schemes with respect to factors such as gender (Gneezy et al. 2003; Niederle and Vesterlund 2007; Delfgaauw et al. 2013), social preferences (Bandiera et al. 2005), task motivation (Ashraf et al. 2014; Butschek et al. 2021), personality traits (Donato et al. 2017), reciprocal inclination (Englmaier and Leider 2020), job mission (Carpenter and Gong 2016) or prior experience (Manthei et al. 2021). In our study, we show that employers can exploit information about worker heterogeneity and increase the performance effect of incentives through a targeted assignment based on the characteristics of individual workers. Thereby our study also adds to a small literature on targeting incentives based on specific individual preferences, e.g., Andreoni et al. (2022) with respect to time preferences and Becker-Peth et al. (2013) with respect to mental accounting. However, to the best of our knowledge, we are the first to apply machine learning to target a broader set of different incentive schemes based on a wider range of individual characteristics.

Our findings also relate to the literature on sorting into incentive schemes. Several studies have shown that individuals sort themselves based on their preferences when choosing between incentive schemes (Lazear 2000; Banker et al. 2000; Cadsby et al. 2007; Dohmen and Falk 2011; Larkin and Leider 2012; Lourenço 2020). However, while this literature has investigated the worker's own sorting decisions, our analysis is, to the best of our knowledge, the first one that studies the targeted assignment of workers to incentive schemes by predicted productivity gains.

Finally, our study complements the growing literature that utilizes machine learning methods to estimate heterogeneous treatment effects and subsequently applies these estimates for optimal policy assignment (see Athey and Imbens (2017) and Athey and Imbens (2019) for comprehensive overviews). Numerous studies offer parametric (Imai and Ratkovic 2013) and non-parametric (Athey and Imbens 2016; Wager and Athey 2018; Farrell et al. 2021b) estimators to identify subgroups with high expected treatment effects while accounting for the issue of multiple hypothesis testing. We compared several of these estimators and found that so-called indirect methods tend to work better in our context than direct methods. With an estimated mapping of individual characteristics to treatment effect in hand, optimal policy assignments can be defined (Allcott and Kessler 2019; Davis and Heller 2020; Godinho de Matos et al. 2018; Hirano and Porter 2009; Hitsch and Misra 2018; Kitagawa and Tetenov 2018; Caria et al. 2020; Farrell et al. 2021a). Such estimated policy assignments are typically used to elicit effect heterogeneity ex-post (Kleinberg et al. 2015, 2017). We extend this line of research by using the predicted optimal assignment to target subjects in a second experiment, thus validating the assignment method out-of-sample (for a related approach in the context of personalized pricing see Dubé and Misra (2023)) and providing new evidence on the external validity of within sample estimates of conditional average treatment effects.

The paper proceeds as follows. First, we present the design and results of Experiment 1 in section 2. Then, in section 3, we explain the implemented algorithm and the resulting assignment procedure. In section 4, we report the results of Experiment 2. Section 5 provides further evidence on how the algorithm raised performance, section 6 studies in detail why the algorithm performed less well on newly hired subjects, and section 7 concludes.

2 Experiment 1

2.1 Experimental Design

The first experiment consists of two parts. First, workers are asked to complete a survey to elicit demographics (i.e., age, gender, education level) as well as personality traits (i.e., Big-5) and social and economic preferences (i.e., social comparison, risk preferences, loss aversion, competitiveness, altruism, positive reciprocity).[4] In the second part, workers work on a real-effort task. We use the real-effort task developed by DellaVigna and Pope (2018), in which workers have to repeatedly press the 'a' and 'b' buttons on their keyboards to score points. One point is awarded for each time they correctly press 'a' then 'b'. Workers have ten minutes to score as many points as possible. Prior to receiving their treatment information, workers have the opportunity to test the task for 30 seconds. We ask them to try to score as many points as possible. We use the points workers score in this test as a proxy for their ability in this type of tasks.[5] After the test phase and a short waiting screen, workers receive information on their treatment.

Workers are randomly allocated to one of six treatments or a control group. Table 1 displays the exact wording of the treatment instructions. Three of these treatments replicate treatments implemented by DellaVigna and Pope (2018) with adapted payment amounts.[6] One of these treatments (*PfP*) is a piece-rate scheme. The other two treatments require the participants to reach a specific goal to receive a bonus and are framed as a gain (*Bonus Gain*) or loss (*Bonus Loss*), respectively. Additionally, three treatments are similar to the ones by DellaVigna and Pope (2018) but are adjusted to make them more comparable to the other three treatments in terms of payments, bonus reached, and guidance on how many points to reach. In particular, we include a gift treatment, where workers receive a bonus without any requirements but are asked to try to reach a specific goal (*Gift & Goal*). Furthermore, we add a treatment which combines a piece-rate for the participants themselves with a performance-contingent donation to charity (*Social PfP*), and a competitive treatment where payments are based on the percentile reached (*Real-time Rank Feedback*). The control group received a fixed wage.

[4] We took the items used to elicit characteristics from the following sources: Big-5 (Benet-Martínez and John 1998, John et al. 1991, John et al. 2008, Rammstedt and John 2007), risk preferences (Falk et al. 2022, 2018), loss aversion (Gächter et al. 2022), competitiveness (Fallucchi et al. 2020), social comparison (Gibbons and Buunk 1999), altruism (Falk et al. 2022, 2018), positive reciprocity (Falk et al. 2022, 2018). Note that participants cannot skip questions, but they can withdraw from the study at any time.

[5] The ability proxy explains a large part of the performance variance in the task (adj. R-squared = 0.167)

[6] We adjusted the payments so that they fitted the different fixed wage we have due to the inclusion of the survey.

Table 1: Treatments

Treatment	Bonus Details
Pay for Performance (PfP)	\$0.05 for every 100 points.
Bonus Gain	\$1 if the score is at least 2000 points.
Gift & Goal	\$1 with the plea to try to score at least 2,000 points.
Bonus Loss	\$1 unless the score is lower than 2,000 points.
Real-time Rank Feedback	\$0.02 times the percentage of former participants who performed worse.
Social PfP	\$0.03 for every 100 points. Plus \$0.02 that go to Doctors Without Borders for every 100 points.
Control	Payment is unaffected.

During the real-effort task, workers see a timer showing the time until the end of the ten minutes, as well as information on the points they have already scored and their current bonus. After completing the task, workers receive information on their total payment and a completion code that they need to submit in order to receive payment.[7]

2.2 Experimental Procedure

We implemented the experiment using oTree (Chen et al. 2016). Workers are invited via MTurk.[8] As common on MTurk, we explicitly advertised our study as an academic study.

Before enrolling in the task, workers are provided with a brief description of the task (complete a survey and a working task) as well as with the technical requirements (a physical keyboard) and guaranteed payment upon successful submission (\$1 flat-pay + \$1.50 guaranteed minimum bonus[9]). Furthermore, they were asked for their consent to participate in the study, from which they know they can withdraw at any time.

[7] See Online Appendix A.3 for screenshots of all instructions.

[8] Evidence suggests that MTurk findings are generally similar to findings in laboratory or field settings (Horton et al. 2011; Farrell et al. 2017; Snowberg and Yariv 2021).

[9] Workers received the guaranteed minimum bonus of \$1.50 for completing the survey. Additional bonuses could be earned in the real-effort task. Please note that workers in the control group also received an additional bonus of \$1 at the end of the study in order to provide them with a reasonably high payment for their participation in the study.

The experiment ran for 2.5 weeks in September 2021. We required workers to be located in the US.[10] In total, 6,649 workers submitted the task for payment. Based on pre-registered criteria[11] we excluded 584 workers resulting in a final sample consisting of 6,065 workers.[12]

The average duration of the experiment was around 20 minutes (median duration around 18 minutes), and the mean payoff was $3.35 ($10.12 per hour; $11.42 per hour median). The mean age in the sample was 39 years, 46.4% of the sample indicated that they were female, 76.3% had at least a college degree. Similar to DellaVigna and Pope (2018) our MTurk sample over-represents somewhat younger and higher educated groups in the U.S. population. In addition, men are somewhat over-represented. Descriptive statistics are shown in Table A1 in the Online Appendix.

The following stratified randomization procedure was applied to achieve balanced sampling into the treatments: Strata were constructed based on the entry time of the workers to the study, i.e. the first seven workers to click on the experiment link and thus enter the study belong to one stratum, the seven workers entering afterwards belong to another stratum, and so on. Within each stratum, treatments 1 to 7 were assigned in a random order such that in each stratum each treatment was assigned once.

2.3 Results of Experiment 1

Figure 1 displays the key results from Experiment 1. All treatments increase performance significantly above the level achieved by the fixed-wage control group ($p < 0.001$). The *Bonus Loss* and *Real-time Rank Feedback* treatment lead to marginally significantly higher performance than the *Social PfP* treatment and significantly higher performance than the *Gift & Goal* treatment.[13]

[10] Further requirements were an approval rate of at least 90% as well as at least 50 approvals. We decided to set requirements relatively low compared to other studies because our working task is not complex, and we were aiming for a large sample size.

[11] As pre-registered, the final sample excludes workers who: (1) do not complete the MTurk task within 90 minutes of starting, (2) are not approved; (3) do not score at least one point, (4) scored 4000 or more points (since this would indicate cheating), or (5) scored 400 or more points in 1 minute (since this would indicate cheating) Restrictions (2)-(4) are the same as in DellaVigna and Pope (2018). Restriction (1) is similar to the restriction in DellaVigna and Pope (2018), however, the maximum completion time is longer due to the survey included in our study. Restriction (5) is equivalent to restriction (4) broken down to individual minutes for which we collected data as well.

[12] The number of workers in the final sample were in *Pay for Performance (PfP)* 879 workers, in *Bonus Gain* 865 subjects, in *Gift & Goal* 875 workers, in *Bonus Loss* 848 workers, in *Real-time Rank Feedback* 874 workers, in *Social PfP* 845 workers, and in *Control* 879 workers. The smaller sample sizes in *Bonus Loss* and *Bonus Gain* mainly come from a larger share of workers which was excluded based on scoring an amount of points that may indicate cheating. The smaller sample size in *Social PfP* mainly comes from more workers withdrawing from the study in this treatment.

[13] This observation is similar to DellaVigna and Pope (2018), where the gift-exchange incentive scheme induced the smallest performance gains. This is also consistent with the results in DellaVigna et al. (2022) who find that MTurk workers receiving a monetary gift increase performance above the level of no incentive but less than with any level of piece rate incentive.

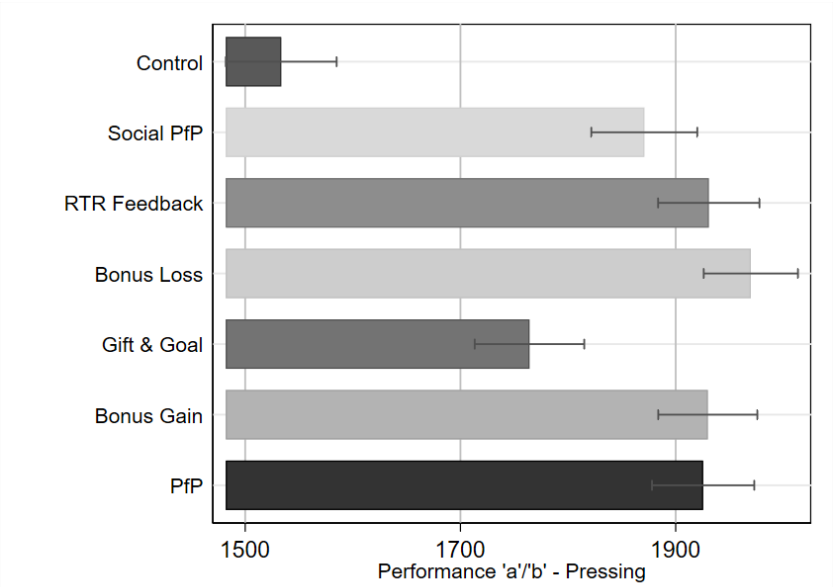


Figure 1: Results of Experiment 1

Note: This figure shows the mean worker performance in Experiment 1 by treatment group. Treatments are described in Table 1. Performance is measured by the number of points scored in the 'a'/'b' - pressing task. Horizontal lines correspond to the 95% confidence interval. For corresponding regression results, see Table A2 in the Online Appendix.

3 Heterogeneity and the Assignment of Incentive Schemes

To assess treatment effect heterogeneity, we first estimate conditional average treatment effects (CATEs) as defined in equation (1).

$$CATE = E[y_i(1) - y_i(0)|X = x] = \tau(x) \tag{1}$$

The CATE is thus the expected difference between the outcome for the individual under treatment $y_i(1)$ and under no treatment $y_i(0)$, conditional on the same characteristics x. If there exists no heterogeneity in the treatment effects, the CATEs do not differ among individuals and coincide with the average treatment effect.

We compared several recently advanced algorithms that combine machine learning and modern causal inference to estimate CATEs. Since algorithms differ in how CATEs are estimated, and there is no a priori guidance as to which algorithm will perform best in our context, we initially employ Causal Forests (Wager and Athey 2018), Causal Nets (Farrell et al. 2021b), Indirect Random Forests (Breiman 2001; Foster et al. 2011) as well as a Doubly Robust approach (Chernozhukov et al. 2018).[14]

Each algorithm then implies a specific optimal assignment policy, which is a mapping from the elicited characteristics to a specific scheme. The scheme assigned to a worker by this mapping is the one that is predicted to yield the highest CATE for this particular worker given the vector of the worker's survey responses.

[14]For the implementation of the Causal Forest and the Doubly Robust approach, we used the EconML python package (Battocchi et al. 2019). For the implementation of the Causal Net, we used the causal_nets python package (https://github.com/PopovicMilica/causal_nets). For the Indirect Random Forest, we used the scikit-learn python package (Pedregosa et al. 2011). We tuned the respective hyperparameters using cross-validation.

To select the best-performing algorithm in our context, we analyzed the results of Experiment 1 using the method in Hitsch and Misra (2018). That is, we train each algorithm on parts of the sample and predict CATEs out-of-sample using cross-validation. We then compared the performance of the implied assignment policies for those out-of-sample observations for which the predicted best assignment coincided with the random assignment in Experiment 1. The performance estimate for that subset of observations is used to select the algorithm with best expected performance on new observations.

Using this evaluation procedure, we found that the indirect random forest approach yielded the highest performance.[15] Based on this finding, we proceed to use the indirect random forest approach to estimate CATEs in the remainder of the study.[16]

The indirect random forest approach involves two steps. We followed the two steps for each of our treatments separately. In step 1, we trained two random forests, one to predict the effort of the treatment group using the personal characteristics elicited by the survey as features, and one to predict the effort of the control group using the same features. Using the estimated models, we predicted the missing counterfactual effort for individuals in each of the two groups. The difference between observed effort and estimated counterfactual effort serves as our initial CATE estimate. In step 2, we used another random forest to model the initial CATE estimates as a function of individual characteristics elicited in the survey.

In addition to standard tuning of algorithm-specific hyperparameters, we also determined the optimal subset of incentive schemes to be implemented in Experiment 2 by maximizing the predicted treatment effect. Using the same method as for the algorithm selection, we compared the performance of the algorithm when restricting the number of potential incentive schemes[17] or after excluding some of the individual characteristics which did not have much predictive power. As a result of this analyses, we restricted the incentive set to *Bonus Loss*, *Real-time Rank Feedback* and *Social PfP*, and did not include a measure of loss aversion and only one of two risk aversion measures as features.[18]

To assess the quality of the algorithmic assignment, we conducted the following exercise: for each group of workers with the same predicted assignment, we compared their performance across the incentive schemes to which they were actually assigned to in Experiment 1. Table 2 shows results. In column (1), we restricted the sample to workers that the algorithm would have assigned to the *Bonus Loss* scheme. Looking at the performance of those workers across the actually assigned schemes, we observed the highest performance gain for the workers that

[15] Table A3 in the Online Appendix shows the performance for each of the algorithms.

[16] While indirect random forests turn out to be the best approach in our setting, other algorithms have been shown to perform well in other contexts (Hitsch and Misra 2018; Farrell et al. 2021b).

[17] We compare the algorithm performance for each combination of three or more incentive schemes. For each combination, we compute the expected performance on those observations for which predicted treatment and actual treatment in Experiment 1 coincide, and we pick the combination that yields the highest expected performance.

[18] That loss aversion does not have much predictive power for performance under the *Bonus Loss* scheme may appear surprising, but is in line with previous evidence showing little predictive power of measures of loss aversion for the endowment effect (Chapman et al. 2017), the decision to accept a loss-framed bonus scheme (De Quidt 2018), or the performance under loss- and gain-framed incentives (Imas et al. 2017). Table A4 in the Online Appendix shows and Figures A1-A3 in the Online Appendix plot the feature importances for the remaining features. Across all trained models, age and altruism are typically among the most relevant predictor variables. Otherwise, the most predictive variables vary by incentive scheme. Overall, we see that demographic features, preference features as well as personality trait features are among the most predictive variables. We do not see that one group is more predictive than the others.

were actually assigned to the *Bonus Loss* scheme. Workers assigned to other schemes also displayed higher performance than the control group but the improvement is much smaller. Similarly, in columns (2) and (3) which restricted the sample to workers that the algorithm would have assigned to *RTR Feedback* or *Social PfP*, respectively, we observed the highest performance increase for those actually assigned to *RTR Feedback* (column (2)) or *Social PfP* (column (3)). Treatment effect differences are significant across all columns with slightly higher standard errors in column (3) due to the smaller sample.[19]

Table 2: Sub-Sample Analysis - Experiment 1

	$log(Performance)_i$		
	Predicted Bonus Loss (1)	Predicted RTR Feedback (2)	Predicted Social PfP (3)
$Bonus\ Loss_i$	0.449***	0.390***	0.312**
	(0.065)	(0.082)	(0.140)
$RTR\ Feedback_i$	0.195***	0.584***	0.384***
	(0.067)	(0.066)	(0.111)
$Social\ PfP_i$	0.196**	0.384***	0.524***
	(0.085)	(0.083)	(0.104)
p-value $Bonus\ Loss_i$=$RTR\ Feedback_i$	0.000	0.005	0.530
p-value $Bonus\ Loss_i$=$Social\ PfP_i$	0.001	0.907	0.081
p-value $RTR\ Feedback_i$=$Social\ PfP_i$	0.995	0.000	0.079
Observations	1,442	1,552	452
Adjusted R-squared	0.149	0.142	0.183

Note: In this table, we report the results of regressions of log(Performance) on treatment dummies for three separate treatments. The sample is restricted to participants in one of the three treatment groups or the control group. The sample is split into sub-samples based on their predicted best treatment using the algorithm trained for Experiment 2. In column (1), the sample is restricted to participants for whom the predicted best treatment is *Bonus Loss*. In column (2) and column (3), the sample is restricted to participants for whom the predicted best treatment is *RTR Feedback* and *Social PfP*, respectively. We include batch fixed effects and an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30-second test phase before they are assigned to a specific treatment. Standard errors are clustered at the batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The importance of specific traits for assigning different schemes can be illustrated in partial dependence plots. For each estimated treatment algorithm, we can, for instance, depict how a change in a particular covariate affects the predicted performance.[20] We then subtract the change in predicted performance for the *Bonus Loss* treatment from the change in predicted performance for the *RTR Feedback* treatment (or *Social PfP* treatment) to get a sense of the range of values of a particular covariate for which predicted treatment effects are higher in the *RTR Feedback* (or *Social PfP*) scheme vis-a-vis the *Bonus Loss* scheme.

[19] Panel (a) in Figure A4 in the Online Appendix plots for each treatment the predicted performance against actual performance. For each treatment, comparisons between actual and predicted performance are close to the 45 degree line, suggesting that algorithmic performance is accurate.

[20] In other words, we calculate the partial dependence of the prediction on changes in a particular covariate keeping all other covariates fixed. See, for example, chapter 10 of Hastie et al. (2009) for details.

Figure 2 shows an example.[21] The upper panels, for instance, illustrate that younger individuals and those with a lower score on altruism are more likely to be assigned to *RTR Feedback* rather than the *Bonus Loss* treatment. Similarly, the lower panels illustrate that younger individuals and those with a higher score on altruism are more likely to be assigned to the *Social PfP* treatment rather than to the *Bonus Loss* treatment.[22]

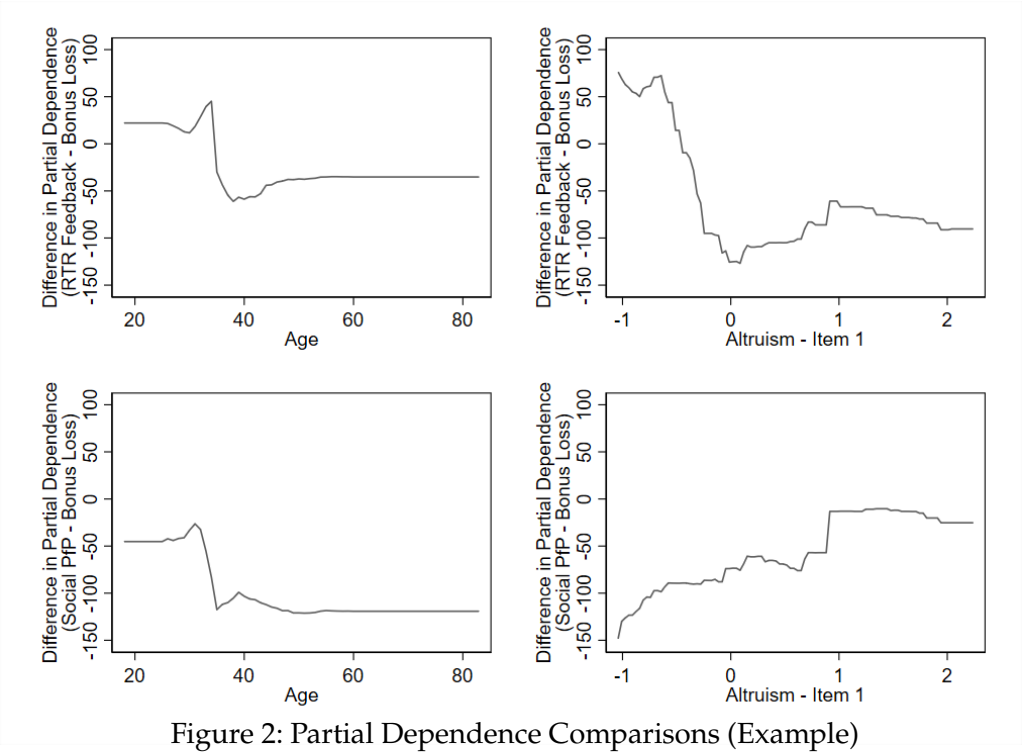


Figure 2: Partial Dependence Comparisons (Example)

Note: This figure shows the difference in partial dependence for the features age and one of our measures of altruism between the *RTR Feedback* scheme and the *Bonus Loss* scheme (i.e. the incentive scheme with the highest point estimate in the first experiment), as well as between the *Social PfP* scheme and the *Bonus Loss* scheme. The altruism item is z-scored. The construction of the plots is described in detail at the end of section 3. See Figure A5 and Figure A6 in the Online Appendix for further partial dependence plots.

4 Experiment 2

4.1 Experimental Design

In Experiment 2, we first elicit workers' characteristics and provide instructions following the same protocol as in Experiment 1.

[21] See Figures A5 and A6 in the Online Appendix for a full set of partial dependence comparisons.

[22] Interestingly, being younger or more altruistic does not by itself lead to assignment to *Social PfP* (note that the difference in predicted performance is always negative), but the patterns suggest that being younger or more altruistic makes such an assignment more likely. Nevertheless, changes in additional features are necessary to change the assignment from bonus-loss to *Social PfP*, which cannot be reflected in the simple ceteris paribus comparison of Figure 2.

Following the survey, workers again receive instructions on the 'a/b'-pressing task and have 30 seconds to test it. We continue to use the points scored in this test phase as a proxy for workers' ability in this task. After the test (and as in Experiment 1), all participants see a waiting screen for 20 seconds, during which participants in the *Algorithm* treatment are assigned to their incentive schemes. After the waiting screen, all workers receive the instructions for the real-effort task, and additional information on their respective incentive scheme. Workers are assigned to one of two treatments or the control group.

In the *Best ATE* treatment, workers are assigned to the incentive scheme with the highest point estimate in Experiment 1, which is the *Bonus Loss* scheme.[23] In the *Algorithm* treatment, workers are assigned to an incentive scheme based on the following procedure: The trained algorithms predict the CATEs for each individual for each incentive scheme based on their elicited characteristics, and they are assigned the treatment with the highest predicted CATE. In the *Control* group, workers receive a fixed wage.

During the real-effort task, workers see a timer showing the time until the end of the ten minutes. Furthermore, they can see how many points they have already scored and how large their current bonus is. After the end of the task, workers receive information on their total payment as well as the completion code, which they need in order to submit the task for payment.

4.2 Experimental Procedure

The procedure of Experiment 2 was similar to that of Experiment 1. The experiment ran over a period of three weeks in November 2021. We again required workers to be located in the US.[24] During the first two weeks, we recruited only workers who had not taken part in Experiment 1. After that, we dropped this restriction as the active pool of new workers on MTurk was exhausted and thus Experiment 2 encompasses both newly hired workers as well as workers who had been part of Experiment 1 before. Due to the sequential randomization procedure, treatment shares were balanced in both populations.[25] In total 6,830 workers submitted the task for payment. We again excluded workers based on the same pre-registered criteria as in Experiment 1 resulting in a sample size of 6,378 workers for the analyses.[26] 4,282 were "New Hires", and 2,096 were "Retaker" who retook the study after having already completed Experiment 1. The sample size of Experiment 2 is based on a power analysis conducted after the first. Specifically, we used the method in Hitsch and Misra (2018) to predict the expected

[23]Note that this choice is a natural benchmark for a risk-neutral decision-maker aiming to maximize expected performance in a new experiment. In our setting, even a risk-averse decision-maker would often choose the *Bonus Loss* scheme as the performance variance is very similar across all incentive schemes (see Figure 1 or Table A2 in the Online Appendix). While *RTR Feedback* has a smaller variance, it also has lower average performance. Note that if the sample composition in a new experiment differs and if such differences were correlated with performance across incentive schemes, a decision-maker might choose a benchmark more suitable for the different sample compositions. This would require reliable knowledge or anticipation of sample differences between existing and new observations. Since both the *Algorithm* treatment and the *Best ATE* treatment are based on data from Experiment 1, we consider the comparison the fairest and most natural given what is known to a decision-maker before running Experiment 2.

[24]Further requirements were an approval rate of at least 90% as well as at least 50 approvals. We decided for these comparable to other studies rather low requirements as our task was not complex, and we were aiming for a large sample size

[25]See Table A5 in the Online Appendix for the balance test results. We analyze the treatment effects in these subpopulations in detail below. See Stewart et al. (2015) for an estimation of the restricted size of the active MTurk population.

[26]The final sample size consists of the following number of workers in the treatments: In *Best ATE* 3,088 workers, in *Algorithm* 3,060 workers, in *Control* 230 workers.

treatment effect of the *Algorithm* treatment in Experiment 2. Based on this predicted effect size, we performed a power analysis for the comparison of the *Algorithm* and *Best ATE* treatments, and determined a sample size of 6,200 workers (3,000 for each treatment group and 200 for the control group).

The average duration of the experiment was around 19 minutes (median duration around 17 minutes) and mean payoff was \$3.33 (\$10.27 per hour; \$11.48 per hour median).[27] 49.3% of workers in Experiment 2 identified as female. 73% had at least a college degree and the mean age was around 39 years old. Once again, our MTurk sample somewhat over-represents younger and more educated groups of the U.S. population. Descriptive statistics are shown in Table A1 in the Online Appendix.

The assignment of the participants to the treatments was determined as follows. First, workers were randomly assigned either to the first control group (i.e., no incentive scheme) or to receive an incentive scheme.[28] For workers who received an incentive scheme, we constructed strata based on the entry time to the study, i.e., the first two workers to click on the link and enter the study belonged to one stratum, the two workers entering afterwards belonged to another stratum and so on. Within these strata, we randomly assigned one individual to the on average best performing treatment in Experiment 1, and another individual was assigned to the treatment suggested by the algorithm.

4.3 Results Experiment 2

While the algorithm still assigned about 39.25% of the subjects to the *Bonus Loss* scheme, a higher share of about 48.01% was assigned to the *Real-time Rank Feedback* condition and a smaller share of 12.75% to the *Social PfP* scheme.[29]

Investigating whether, and if yes, to what extent the algorithmic assignment of the scheme can improve performance, Column (1) in Table 3 shows a regression of the log performance in Experiment 2 on two treatment dummies. The $Best\ ATE_i$ dummy indicates that observation i has been assigned to the treatment where all workers were exposed to the scheme with the highest average treatment effect (the *Bonus Loss* scheme).[30] The $Algorithm_i$ dummy indicates an observation from the treatment where the assignment is based on the algorithm. In Columns

[27] As in Experiment 1, participants in the control group receive an additional \$1 bonus at the end of the study in order to reasonably compensate them for their participation.

[28] The probability of being assigned to no incentive scheme was adjusted to around 3% such that we would get the preregistered sample size of around 3,000 workers for each incentive treatment and around 200 workers for the control group. We aimed for the smaller sample size in the control group as power analyses showed that this small size was sufficient for high power.

[29] Instead of maximizing expected performance of workers the same procedure could of course also be applied for other objectives such as the minimization of the costs per unit of output or the maximization of profits (for a specific value of a unit of output). See Table A6 in Online Appendix for the summary of the costs per treatment and incentive scheme. The set of schemes selected will depend on the respective objective. For instance, when assigning the schemes with respect to the minimization of unit costs more subjects would be assigned to *Bonus Loss* and *Social PfP* schemes and fewer to the *Real-time Rank Feedback*. Note, however, that whenever the value of output is sufficiently large, the profit maximizing scheme assignment will be identical to the performance maximizing one for which we have opted here.

[30] Table A7 in the Online Appendix shows the results when controlling for retaking as well as the Retakers' Experiment 1 treatment assignment. We find no evidence that treatment assignment in Experiment 1 affects performance in Experiment 2. We also do not find evidence that treatment assignment has an influence on becoming a Retaker (see Table A8 in the Online Appendix). See Table A9 in the Online Appendix for the results using absolute performance.

(2)-(4), we restrict the sample to workers in one of the two treatment groups so that we can directly compare their performance. As Experiment 2 included subjects that had already taken part in Experiment 1 ("Retakers") and newly hired subjects ("New Hires") we split the sample into these two subgroups in columns (3) and (4).

Table 3: Main Results: Effect on Performance

	$log(Performance)_i$			
	All	All	New Hires	Retakers
	(1)	(2)	(3)	(4)
$Algorithm_i$	0.257***	0.043**	0.016	0.097***
	(0.057)	(0.017)	(0.023)	(0.028)
$Best\ ATE_i$	0.214***			
	(0.058)			
p-value $Best\ ATE_i$=$Algorithm_i$	0.013			
Reference Group	Control	Best ATE	Best ATE	Best ATE
Observations	6,377	6,147	4,131	2,015
Adjusted R-squared	0.112	0.110	0.099	0.132

Note: In this table, we report the results of regressions of log(Performance) on treatment dummies for the *Best ATE* (Bonus Loss) treatment as well as the *Algorithm* treatment. In columns (2), (3) and (4), we exclude the control group so that *Best ATE* is the reference group for the *Algorithm* dummy. In column (3), we restrict the sample to the newly hired workers in Experiment 2. In column (4), we restrict the sample to the workers who have already taken part in Experiment 1 (Retakers). We include batch fixed effects as well as an ability proxy as control. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Performance is measured as 'a/b'-presses in a 10 minute time window. Standard errors are clustered at the batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The first key result is that over the whole sample, the algorithmically assigned scheme indeed significantly outperforms the average best scheme from Experiment 1: As Table 3 shows, the *Best ATE* treatment raises performance above the level of the fixed wage control group by about 23.9%. The effect of the *Algorithm* treatment is 29.3%. The targeted assignment of incentive schemes thus significantly increases the overall incentive effect by 5.4 percentage points or 22.5% ($p = 0.013$). This corresponds to a more than 4% increase in performance compared to the group working under the single best incentive scheme in Experiment 1.

However, as columns (3) and (4) show, this effect is driven primarily by subjects that had been part of Experiment 1 (column (4)). In this group, the algorithm outperforms the average best scheme by about 10%. In the sample of newly hired workers (column (3) the respective point estimate is only about 1.6% and not significantly different from zero. Hence, a second main observation is that the targeted assignment of incentives succeeded only when deployed to workers on which the respective algorithm was trained, a finding we will explore in detail in section 6.

5 How Did the Algorithm Raise Performance?

5.1 Effects by Incentive Scheme

In a next step, we decompose the overall effect into the effects obtained by assigning workers to the specific scheme that is predicted to be superior to the *Bonus Loss* scheme. To do that, we split the complete sample from Experiment 2 into sub-samples by the respective scheme assigned by the algorithm based on a person's characteristics.[31] Within each of these sub-samples, we estimate the average treatment effect of the respective scheme assigned by the algorithm in comparison to the *Best ATE* treatment (i.e. the *Bonus Loss* scheme). Results are displayed in Table 4.[32]

The first sub-sample comprises all subjects from the three treatments for which the algorithm predicted that their performance is highest under the *Bonus Loss* scheme. Note that here the *Best ATE* and the *Algorithm* treatments implement exactly the same scheme on a sub-sample selected by exactly the same procedure and thus both point estimates have the same magnitude.

The second sub-sample comprises all subjects which the algorithm would assign to the *Real-time Rank Feedback* scheme. In this sub-sample, the assignment by the algorithm to the *Real-time Rank Feedback* scheme raises performance by more than 7% compared to the performance under the *Bonus Loss* scheme.

Table 4: Effects in Sub-Samples

	$log(Performance)_i$		
	Predicted Bonus Loss (1)	Predicted RTR Feedback (2)	Predicted Social PfP (3)
$Algorithm_i$	0.008	0.069***	0.018
	(0.043)	(0.025)	(0.066)
Reference Group	Best ATE	Best ATE	Best ATE
Observations	2,432	2,906	805
Adjusted R-squared	0.100	0.102	0.136

Note: In this table, we report the results of regressions of log(Performance) on an *Algorithm* treatment dummy in sub-samples split by the predicted best treatment. We exclude the control group so that *Best ATE* is the reference group for the *Algorithm* dummy. Column (1) presents the results for the sub-sample of all participants (regardless of their actual assignment) for which the *Bonus Loss* was predicted to be the best incentive scheme based on their individual characteristics. Column (2) and (3) present the results for the sub-sample of all workers (regardless of their actual assignment) for which the *Real-time Rank Feedback* and *Social PfP* was predicted to be the best incentive scheme based on their individual characteristics, respectively. We further include batch fixed effects and an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The last sub-sample comprises subjects predicted to achieve the highest performance in *Social PfP*. Within this group, the *Social PfP* (which led to a weaker performance than loss in Experiment 1) catches up to the *Bonus Loss* scheme. The respective point estimate is positive but insignificant.

[31] That is the observations from the *Best ATE* treatment are allocated to the subsample associated to the scheme that the algorithm would have assigned them to.

[32] See Table A10 in the Online Appendix for the results split by New Hires and Retakers.

Hence, the overall effect is driven by participants, which the algorithm assigns to the *Real-time Rank Feedback*. These participants show a large increase in performance when assigned to their predictably best treatment.

5.2 Assignment Group Characteristics

As the algorithm utilizes the abundant information contained in the different patterns of survey response behavior and potentially complex interaction structures, it is not possible to depict the specific functional form employed. Nevertheless, it is instructive to examine which of the measured characteristics are directly associated with the likelihood of a being person assigned to a specific scheme.

To illustrate this point, we estimate simple logistic regressions of a dummy indicating the assignment to a specific scheme on demographic characteristics as well as key aggregated preference and personality measures. The results are reported in Table 5.[33]

Note that all the measures of preferences and personality traits have been standardized to compare the magnitudes of the respective regression coefficients. Several features stand out. Older workers and females are significantly more likely to be assigned to the *Bonus Loss* scheme. The latter is in line with previous findings that women tend to be more loss averse than men (e.g., Rau 2014; Andersson et al. 2016), which would imply that they exert more effort to avoid a loss.[34] It is also in line with previous research that has shown that women perform less well under competitive incentives (see e.g. Gneezy et al. 2003), women are less likely to be assigned to the *Real-time Rank Feedback* scheme. However, somewhat surprisingly, women are also less likely to be assigned to the *Social PfP*.[35]

Among the trait and personality measures, we observe the most pronounced differences with respect to altruism. As per straightforward reasoning, subjects with more altruistic tendencies are significantly more likely to be assigned to the *Social PfP* scheme and less likely to be assigned to *Real-time Rank Feedback*. Moreover, positive reciprocity, agreeableness, and extraversion, all of which are associated with prosocial traits, are positively associated with the probability of being assigned to *Social PfP*.

Unexpectedly, our survey measure of competitiveness is associated with a significantly lower likelihood of being assigned to the competitive *Real-time Rank Feedback* scheme and a higher likelihood to work under the *Bonus Loss* scheme.[36] We also find that more risk-averse individuals are more frequently assigned to *Real-time Rank Feedback* and less often to the *Bonus Loss* scheme.[37]

[33] Figure A7 in the Online Appendix plots the averages of each characteristic in the three groups. Also, Table A11 in the Online Appendix shows the results for the group of New Hires separately.

[34] Note that we also had included a survey measure of loss aversion in our initial survey, but this measure has turned out not to be predictive for the conditional average treatment effects and thus was dropped in the assignment procedure for Experiment 2.

[35] While some papers such as Tonin and Vlassopoulos (2010) and Drouvelis and Rigdon (2022) find that women are more motivated through social incentives than men, Tonin and Vlassopoulos (2015) and Imas (2014) do not find significant gender differences in response to social incentives.

[36] A potential interpretation is that a distaste for competition may not automatically imply a lower performance under competition. Grund and Sliwka (2005), for instance, show in a formal model that inequality aversion raises performance in a tournament (as agents work harder to avoid disadvantageous inequality) but at the same time lowers the preference to join a tournament.

[37] Skaperdas and Gan (1995) show that it can be rational for risk averse agents to work harder in contests than the less risk averse.

Table 5: Group Characteristics (Logit)

	Predicted Bonus Loss$_i$ (1)	*Predicted RTR Feedback*$_i$ (2)	*Predicted Social PfP*$_i$ (3)
Age$_i$	0.078***	-0.018***	-0.262***
	(0.005)	(0.003)	(0.014)
Female$_i$	1.253***	-0.949***	-1.365***
	(0.107)	(0.081)	(0.172)
Some College$_i$	-0.018	0.114	-0.679**
	(0.146)	(0.154)	(0.324)
Bachelor's Degree or more$_i$	0.099	-0.241	0.575**
	(0.146)	(0.150)	(0.236)
Ability Proxy$_i$	-0.078**	0.183***	-0.107**
	(0.033)	(0.040)	(0.044)
Conscientiousness$_i$	0.389***	-0.224***	-0.590***
	(0.058)	(0.054)	(0.070)
Openness$_i$	-0.277***	0.239***	0.048
	(0.040)	(0.050)	(0.084)
Emotional Stability$_i$	0.052	-0.041	-0.163*
	(0.057)	(0.065)	(0.085)
Agreeableness$_i$	-0.012	0.031	0.184***
	(0.052)	(0.057)	(0.068)
Extraversion$_i$	0.386***	-0.546***	0.557***
	(0.041)	(0.048)	(0.081)
Altruism$_i$	0.673***	-1.931***	2.091***
	(0.065)	(0.092)	(0.120)
Positive Reciprocity$_i$	-0.526***	0.417***	0.343***
	(0.053)	(0.063)	(0.091)
Competitiveness$_i$	1.079***	-1.101***	-0.128
	(0.057)	(0.046)	(0.107)
Social Comparison$_i$	0.135**	-0.004	-0.361***
	(0.065)	(0.061)	(0.101)
Risk Aversion$_i$	-0.672***	0.621***	0.020
	(0.049)	(0.056)	(0.069)
Observations	6,378	6,378	6,378
Pseudo R-squared	0.319	0.441	0.463

Note: In this table, we report the results of a logistic regression of a dummy of having *Bonus Loss* (column (1)), *RTR Feedback* (column (2)), or *Social PfP* (column (3)) as predicted best incentive scheme on the features the algorithm uses for assignment. With the exception of age (continuous), female (binary), some college (binary) and bachelor's degree or more (binary) all variables are standardized. For all characteristics for which we used more than one item as a feature, we built a summative scale (i.e. for the big-5, altruism, positive reciprocity, competitiveness and social comparison). Standard errors are clustered at the batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Finally, upon close examination of Tables 4 and 5, it appears that payouts differ by demographic characteristics, in particular by gender and age. Table A12 in the Online Appendix shows, for instance, that female participants benefit less on average when algorithmically assigned to incentive schemes (although both male and female participants benefit on average). Results for pay disparity by age are similar; workers benefit from algorithmic assignment regardless of age, but younger workers tend to benefit more (see Table A13 in the online appendix). The gender pay gap is entirely explained by individuals assigned to the *Real-time Rank Feedback* scheme, which appears to benefit male participants more than female participants (female participants actually earn a bit more on average in the *Social PfP* scheme). This finding might raise concerns about algorithmic discrimination and fairness (Rambachan et al. 2020). For example, a decision-maker might prefer an assignment that yields equal payouts regardless of gender (or age). Note that algorithms such as ours can be used to target such objectives specifically. In particular, while our objective was to find an assignment that maximizes effort, it is easily conceivable to add fairness considerations as potential constraints to the optimization problem in practice.

6 Why Did the Algorithmic Assignment Perform Better on the "Retakers"?

Our average treatment effect is mainly driven by subjects who had already participated in Experiment 1 (referred to as "Retakers"). It is thus important to investigate why the algorithmic assignment was less effective for the subjects who did not participate in Experiment 1 (referred to as "New Hires"). There are several different possible explanations which we will disentangle below. First, the sample characteristics may differ too much between the training sample and the sample of New Hires, making it more challenging for the algorithm to assign the best incentive scheme for the New Hires (an issue discussed in the Machine Learning literature as "covariate or dataset shift"). Secondly, it is conceivable that subjects who have a larger propensity for future interactions with employers on the platform inherently differ from one-shot participants. For example, one-shot participants might be less attentive to instructions, less thorough in filling out the survey, or less consistently acting based on their traits when performing the work task. Consequently, it is likely that the algorithm performs worse on these one-shot participants in general (we will refer to this as "propensity for future interactions"). Thirdly, and related to the previous point, if New Hires give less consistent survey responses than Retakers, noise in the survey responses may naturally have limited the transferability of the learned patterns due to measurement error (we will refer to this as "measurement error"). Lastly, the difference may be due to the fact that Retakers were more experienced with the task or the survey (we will refer to this as "experience").

6.1 Covariate Shift in Observables

A first potential explanation for the lack of effectiveness of the assignment among New Hires is that they differ too strongly in their observable characteristics. Indeed, when we compare descriptive statistics in key observed traits, we see some sizeable differences, as shown in Table A14 in the Online Appendix. For instance, New Hires are significantly younger, less educated, and the share of women is larger among them. As is well known in the literature on machine learning (Quinonero-Candela et al. 2008; Moreno-Torres et al. 2012; Ovadia et al. 2019), such a covariate shift can limit the power of algorithms trained on one sample to make precise predictions for other related samples.

To assess the role of such a shift in covariates, we proceed as follows: We first generate a measure of similarity of a New Hire with the training sample. To do so, we pool the data of all subjects in Experiment 1 and the New Hires from Experiment 2. We then train a Random Forest based on all observable covariates to predict the likelihood that an observation was part of Experiment 1.[38] This predicted probability to be an observation from Experiment 1 rather than a New Hire serves as a measure of similarity to the training set.

To assess whether covariate similarity indeed matters for the usefulness of the algorithm in assigning incentives, we first sort New Hires into quartiles based on their estimated similarity to Experiment 1.[39] Within the sample of New Hires, we then regress log performance on the *Algorithm* dummy interacted with dummies indicating quartiles of the respective observations' similarity to the training sample.

If a shift in observable covariates indeed explains why the algorithmic assignment performed worse on the New Hires, we should see (i) a treatment effect among those New Hires who are most similar to the training sample with respect to their observable covariates and (ii) reduced treatment effects for less similar subjects. But in fact, our regression results reported in column (1) of Table 6 do not provide evidence for either of the two effects.[40] Hence, a shift in observable covariates is unlikely to explain the lack of an effect among the New Hires.

6.2 Propensity for Future Interactions

As discussed above, there might also be an inherent difference in (potentially unobserved) traits between subjects who have a high propensity to repeatedly work on MTurk and one-shot participants. It is conceivable that subjects who have a higher propensity to repeatedly offer their services on the platform are more attentive when reading the instructions, filling out the survey, or more consistent in their actions when performing the work task. To investigate this channel, we assess this propensity for future interactions and investigate whether the algorithm is more effective in a subsample of the New Hires who have a sufficiently high propensity for future interactions.

[38] To be precise, we performed 100 splits training on one half to make predictions for the other half and vice versa. We use the mean of the respective predictions for the propensity to be an Experiment 1 observation.

[39] That is, we rank the observations by their predicted probability to be in the training set such that observations in Q1 belong to the top quartile and those in Q4 to the bottom quartile.

[40] We applied several alternative approaches to study the role of observable of these sample differences including nearest-neighbor matching, propensity score matching, inverse propensity score weighting, and inference on counterfactual distributions. All these analyses indicate that the heterogeneous treatment effects are not explainable with a shift in observable covariates.

Table 6: Mechanisms - New Hires

	$log(Performance)_i$			
	Training Similarity (1)	Future Interaction (2)	Consistency (3)	Fut. Int. & Consistency (4)
$Algorithm_i$	0.052	0.095***	0.087***	0.132***
	(0.044)	(0.026)	(0.023)	(0.027)
$\times$ *Training Similarity Q2*$_i$	-0.041			
	(0.052)			
$\times$ *Training Similarity Q3*$_i$	-0.047			
	(0.063)			
$\times$ *Training Similarity Q4*$_i$	-0.056			
	(0.051)			
$\times$ *Future Interaction Q2*$_i$		-0.065		-0.053
		(0.048)		(0.046)
$\times$ *Future Interaction Q3*$_i$		-0.089*		-0.070
		(0.046)		(0.045)
$\times$ *Future Interaction Q4*$_i$		-0.167***		-0.127**
		(0.055)		(0.054)
$\times$ *Consistency Q2*$_i$			0.006	0.017
			(0.030)	(0.031)
$\times$ *Consistency Q3*$_i$			-0.128***	-0.110**
			(0.044)	(0.043)
$\times$ *Consistency Q4*$_i$			-0.164***	-0.129**
			(0.061)	(0.062)
Reference Group	Best ATE	Best ATE	Best ATE	Best ATE
Observations	4,131	4,131	4,131	4,131
Adjusted R-squared	0.098	0.100	0.101	0.102

Note: In this table, we report the results of regressions of log(Performance) on a dummy for being in the *Algorithm* treatment. We exclude the control group so that the *Best ATE* treatment group is the reference group for the *Algorithm* dummy. In column (1), we further include the interactions between *Algorithm* and being in the second, third or lowest quartile of the New Hire sample with regards to their predicted similarity with the training sample. We predict the similarity with the training sample using a simple random forest model trained on a pooled sample of Experiment 1 and the New Hires using 2-fold cv with 100 random sample splits and averaging the out-of-sample predictions. In column (2), we further include the interactions between *Algorithm* and being in the second, third or lowest quartile of the New Hire sample with regards to their predicted propensity for future interaction. We predict propensity for future interaction using a simple random forest model trained on the Experiment 1 data including information who became a Retaker later on. In column (3), we further include interactions between *Algorithm* and being in the second, third or lowest quartile regarding consistency. The measure for the consistency of survey answers is the z-scored reversed mean absolute distance between mean answers to originally reversed-coded and normally coded items of the measured characteristics (after reversing the scales so that they are coded in the same direction). In column (4), we include predicted propensity for future interaction as well as consistency quartiles. See Table A15 in the Online Appendix for regressions where training similarity, as well as future interaction, consistency, or both, are included. We include batch fixed effects as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

In order to do so, we train a random forest to predict the likelihood that someone will become a Retaker (i.e. will actually accept the later invitation to Experiment 2) based on the survey responses from Experiment 1. We then apply this algorithm to predict the propensity for future interactions for each New Hire taking part in Experiment 2. This predicted probability serves as a measure of the propensity of a person to register for the same job in a future invitation. Analogous to the above analysis, we again only consider the set of New Hires in Experiment 2 and regress their log performance on a dummy for the *Algorithm* treatment interacting the treatment with the respective quartile of the propensity for future interactions. The respective regression results are shown in column (2) of Table 6.

As the table shows, the algorithm performs quite well even for New Hires – when these New Hires have a sufficiently high propensity for future interactions. In the top quartile of New Hires, the *Algorithm* outperforms the *Best ATE* by more than 9%, which is close to the treatment effect among the Retakers. However, as the propensity decreases, the treatment effect becomes smaller and vanishes for those with low propensity for future interactions.[41] In other words, the treatment effect is larger the higher the propensity of a worker to be similar to a Retaker.[42] This suggests that an inherent difference between Retakers and one-shot participants is at least partly responsible for the difference in treatment effects.

6.3 Measurement Error

The arguments presented above suggest that the (lack of a) propensity for future interactions may be related to the sloppiness with which subjects will fill out the survey, consequently undermining the usefulness of the provided information for the assignment procedure. When there is more measurement error in assessing traits, it becomes more challenging for the algorithm to assign the incentive scheme that maximizes performance.

Therefore, we delve into a more detailed investigation of the role of consistency in survey responses. We generate a consistency measure using the responses to different survey items measuring the same trait. We take advantage of the fact that several of the psychological scales we used include reverse-coded items.[43] To quantify consistency, we calculate the z-scored reversed mean absolute distance between mean answers to originally reversed-coded and normally coded items of the measured characteristics (after reversing the scales so that they are coded in the same direction).

In line with the above conjecture, we find that Retakers indeed exhibit significantly greater consistency in their answering behavior. Table A16 in the Online Appendix shows that consistency is significantly larger for Retakers both when they fill out the survey in Experiment 1 and in Experiment 2. That is, Retakers, on average, provide more consistent answers than one-shot participants (that is, subjects in Experiment 1 who then don't take part in Experiment

[41] The point estimate of the algorithm treatment in Q4 (i.e. the sum of the *Algorithm* coefficient and the respective interaction term) even becomes negative, but is not significantly different from zero.

[42] See Table A17 in the Online Appendix for the treatment effects in subgroups of Retakers and New Hires separately whose propensity for future interaction is above certain thresholds.

[43] This applies to the following traits: conscientiousness, agreeableness, emotional stability, extraversion, openness, and social comparison. For instance, the conscientiousness scale includes the items "I see myself as a person who does a thorough job." and "I see myself as a person who tends to be lazy."

2 and subjects in Experiment 2 who have not taken part in Experiment 1 before). In addition, our measures of the propensity for future interactions and consistency are positively correlated both within the sample of Retakers (correlation coefficient of 0.34) and within the sample of New Hires (0.33).

We can now go one step further to investigate whether the *Algorithm* treatment may outperform the *Best ATE* among the most consistent subjects when we again consider only the sample of New Hires.

We proceed analogously to the previous analyses, by considering the quartile ranking of New Hires with respect to the consistency measure interacting the respective quartile dummies with the treatment coefficient. The regression results are displayed in column (3) of Table 6. As the results show, the *Algorithm* outperforms the *Best ATE* treatment by close to 9% when we only consider the 25% of the most consistent survey respondents among the New Hires. The magnitude of the treatment effect remains fairly similar even when considering the top 50% respondents, but it deteriorates when including the least consistent respondents.[44]

As a complementary analysis, we generate an alternative (and more comprehensive) measure of consistency for all Retakers as these subjects have filled out the survey twice. For these subjects, we can directly measure the test-retest reliability (a standard measure used to assess the reliability of surveys in Psychology) by computing the correlation between individual survey responses from both experiments. As the survey uses different scales for different traits, we use the Spearman-rank correlation coefficient and compute the correlation for each individual subject. The median test-retest reliability is 0.72 (mean: 0.66). When we only use the sample of Retakers, we can interact the treatment effect with the test-retest reliability. The respective regression results are shown in Table A18 in the Online Appendix and again show that the algorithm is substantially more effective when subjects provide more consistent answers.

One interesting implication of this finding is that inconsistency itself does not seem to be indicative of traits that influence incentive effects. This is noteworthy because we trained the algorithm with raw data from all survey items, which may have picked up personality traits that could have revealed itself in more or less consistent responses.

A final key question is whether the propensity for future interactions and the response consistency independently contribute to the understanding of the lower effectiveness of algorithmic assignment among New Hires. To study this, we add both interactions in column (4) of Table 6. Indeed, we find the largest treatment effects within the group of New Hires that are most consistent and most likely to become a Retaker. Here, the increase in performance of the *Algorithm* treatment relative to the *Best ATE* group is about 14%. The results also indicate that while the inclusion of the consistency interactions somewhat weakens the size of the propensity for future hiring interactions, the latter remains sizeable and significant (and vice versa). Summing up, both measures seem to contribute to the core treatment effect heterogeneity.

[44] See Table A19 in the Online Appendix for the treatment effects in subgroups of Retakers and New Hires separately whose consistency is above certain thresholds.

6.4 Experience

A further difference between New Hires and Retakers is that the Retakers have previous experience with the task and the survey when they take part in Experiment 2. As the algorithm had been trained on data from this population when they performed the task for the first time, it appears unlikely that the higher effectiveness of the algorithm among Retakers is due to their experience with the task. However, it is conceivable that the survey is more informative for subjects filling it out for the second time. Hence, it is still worth exploring this difference in more detail.

To evaluate potential experience effects, we investigate how the algorithm would have performed had it been available for Experiment 1. Our data allows us to do this by considering only workers from Experiment 1 that (by chance) ended up in the scheme that the algorithm would actually have assigned based on their survey responses. We can then compare the performance of these workers in Experiment 1 to the performance of workers in the *Bonus Loss* treatment in the same experiment. A corresponding regression is reported in Table A20 in the Online Appendix. The point estimate of this "hypothetical algorithm treatment" is 0.113, which is very close and even slightly larger than the *Algorithm* treatment effect of 0.097 observed in Experiment 2, suggesting that experience, if at all, reduced performance (the difference between these treatment effects is not statistically significant). Therefore, it seems unlikely that the algorithm's better performance among Retakers is due to their experience with either the task or the survey.

To sum up, why did the algorithm not raise performance in the sample of all New Hires? We find that neither experience with the task nor differences in observable covariates between New Hires and the training sample can explain differences in performance. Instead, we find that within the sample of New Hires that i) answer the survey consistently and that ii) are likely to interact on the platform repeatedly (a high propensity for future interactions), treatment effects are of similar magnitude to treatment effects among the sample of Retakers. This suggests that inaccurate measurement of traits for a share of one-shot participants and other unobservable differences between Retakers and one-shot participants at least partly explain the difference in treatment effects.

7 Conclusion

Our study demonstrates that targeted assignment of incentive schemes based on individual worker characteristics can elevate overall worker performance beyond the level achieved by a scheme that performs best on average. Moreover, we show that even unincentivized survey measures of preferences and traits are useful predictors of heterogeneous responses to different incentive schemes.

Our results have several implications for the design of incentive schemes. Organizations may consider to use individual worker characteristics to assign incentive schemes that, in turn, increase workers' performance. Of course we caution that individually targeting different incentives within a team may lead to pay inequity and potential adverse effects resulting from it (see, e.g., Breza et al. 2018). Hence, there will be limitations to apply this approach in traditional

organizations. Yet, the rise of alternative work arrangements (Katz and Krueger 2019), especially the gig economy, opens a particularly suitable field for the assignment of different schemes to different workers due to the independent work environment typical for gig work. Indeed, it has been shown that gig workers can respond to incentives quite differently, even within the same job (Butschek et al. 2021). Yet, our approach can, in principle, also be applied to more traditional organizations if the assignment is done on a team rather than the individual level – provided that teams differ in their underlying characteristics (such as the type of tasks performed). The approach holds the potential not only to raise performance but also to improve employee well-being and satisfaction if supported by appropriately trained algorithms. Finally, a company that can only offer a single incentive scheme to all employees can still benefit from evidence on targeting: Here, such a company could focus on hiring workers that are best motivated by the incentive scheme that the company can offer, e.g. if the company can offer only pay-for-performance incentives, it could hire workers that are deemed the most productive under such a scheme given their elicited preferences and traits.

Looking ahead, future research should investigate the possible difference between workers' own selection into different incentive schemes and the algorithmic assignment in more detail. It is likely that the preferences of workers for certain incentives differ from what is best to increase their performance (see, e.g., Lourenço 2020).[45]

Related, workers may, for instance, be able to manipulate algorithmic decision rules to get assigned to their preferred incentive scheme. If this is a concern, one solution might be to only base the algorithmic decision rule on inputs that are costly or impossible to manipulate (in practice, many firms keep the decision rule hidden). Björkegren et al. (2020) develop an estimator that can take individual incentives to manipulate input data into account when the cost of manipulation can be quantified.

Our results also highlight the limitations of the algorithmic approach. Importantly, measurement error in assessing traits makes it harder for the algorithm to assign optimal incentive schemes. Hence, reliable survey responses are key to successful assignments, urging researchers and practitioners to prioritize the quality of elicited characteristics. As we have shown, the issue of unreliable measurement of worker characteristics appears particularly prevalent for workers who do not have a longer term perspective. Targeted assignment may thus be less effective in those setting. On the other hand, workers interested in continuing interactions are likely to be the most relevant in an actual work/hiring context outside of our study. For such workers, it should be easier to collect more reliable information on traits from actual job testing. From that point of view, our results that use non-incentivized self-reported information for the targeting may even be interpreted as a lower bound on treatment effects that could be measured in realistic work environments.

Adapting the approach to other work environments faces additional challenges. First, the approach requires access to objective performance data, and it may be less powerful when performance can only be assessed subjectively. As it is well-known that subjective performance assessments tend to be biased, an assignment based on maximizing subjective performance

[45] Yet, algorithms may even be useful in settings where workers can self-select into schemes, if, for instance, workers are uncertain about their own preferences and targeted predictions may help to find out individually optimal choices.

assessments may well differ from the assignment that maximizes objective performance.[46] Still it appears worthwhile to investigate proxy measures of objective performance such as project completion, attendance of employees, or customer feedback scores. Second, the relation between relevant worker traits and optimally assigned incentive schemes will likely be different across work tasks with different characteristics. In particular, an important question for future research concerns the transferability of a trained algorithm across tasks, i.e. whether an algorithm trained on a specific task can also be used for optimal assignment in other tasks that differ in some (but not all) dimensions.

Finally, the application of our approach must allow for a sufficiently large number of workers working under different incentive schemes such that heterogeneous patterns can be estimated reliably. Nevertheless, as data availability increases, the utilization of richer sources of information, such as digital footprints, holds promise for maximizing the impact of targeted assignment (Youyou et al. 2015; Azucar et al. 2018). Workers might not always be aware that the data that they consciously and unconsciously provide can be used for such purposes. Balancing the desire of firms to optimally allocate resources with the desire of workers for data privacy will remain a delicate trade-off for years to come.

[46] This is a common problem also in the training of algorithms for personnel selection and hiring: When the algorithm is trained to predict subjectively measured success, it may also replicate the stereotypes of the managers providing the subjective assessments.

References

Allcott, H. and J. B. Kessler (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics 11*(1), 236–76.

Andersson, O., H. J. Holm, J.-R. Tyran, and E. Wengström (2016). Deciding for others reduces loss aversion. *Management Science 62*(1), 29–36.

Andreoni, J., M. Callen, M. Y. Khan, K. Hussain, and C. Sprenger (2022). Using preference estimates to customize incentives: An application to polio vaccination drives in pakistan. *Journal of the European Economic Association*, 1–50.

Armantier, O. and A. Boly (2015). Framing of incentives and effort provision. *International Economic Review 56*(3), 917–938.

Ashraf, N., O. Bandiera, and B. K. Jack (2014). No margin, no mission? a field experiment on incentives for public service delivery. *Journal of Public Economics 120*, 1–17.

Athey, S. and G. W. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*(1), 685–725.

Azucar, D., D. Marengo, and M. Settanni (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences 124*, 150–159.

Bandiera, O., I. Barankay, and I. Rasul (2005). Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics 120*(3), 917–962.

Bandiera, O., I. Barankay, and I. Rasul (2007, 05). Incentives for Managers and Inequality among Workers: Evidence from a Firm-Level Experiment*. *The Quarterly Journal of Economics 122*(2), 729–773.

Bandiera, O., I. Barankay, and I. Rasul (2011, September). Field experiments with firms. *Journal of Economic Perspectives 25*(3), 63–82.

Banker, R. D., S.-Y. Lee, G. Potter, and D. Srinivasan (2000). An empirical analysis of continuing improvements following the implementation of a performance-based compensation plan. *Journal of Accounting and Economics 30*(3), 315–350.

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. https://repository.upenn.edu/bepp_papers/75. Accessed: 2022-03-01.

Battocchi, K., E. Dillon, M. Hei, G. Lewis, P. Oka, M. Oprescu, and V. Syrgkanis (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML. Version 0.x.

Becker-Peth, M., E. Katok, and U. W. Thonemann (2013). Designing buyback contracts for irrational but predictable newsvendors. *Management Science 59*(8), 1800–1816.

Benet-Martínez, V. and O. P. John (1998). Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of Personality and Social Psychology 75*(3), 729.

Björkegren, D., J. E. Blumenstock, and S. Knight (2020). Manipulation-proof machine learning. Working paper, arXiv:2004.03865.

Blanes i Vidal, J. and M. Nossol (2011). Tournaments without prizes: Evidence from personnel records. *Management Science 57*(10), 1721–1736.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breza, E., S. Kaur, and Y. Shamdasani (2018). The morale effects of pay inequality. *The Quarterly Journal of Economics 133*(2), 611–663.

Butschek, S., R. G. Amor, P. Kampkötter, and D. Sliwka (2021). Motivating gig workers–evidence from a field experiment. *Labour Economics 75*, 102105.

Cadsby, C. B., F. Song, and F. Tapon (2007). Sorting and incentive effects of pay for performance: An experimental investigation. *Academy of Management Journal 50*(2), 387–405.

Caria, S., G. Gordon, M. Kasy, S. Quinn, S. Shami, and A. Teytelboym (2020). An adaptive targeted field experiment: Job search assistance for refugees in jordan. Working paper no. 8535, CESifo.

Carpenter, J. and E. Gong (2016). Motivating agents: How much does the mission matter? *Journal of Labor Economics 34*(1), 211–236.

Casas-Arce, P. and F. A. Martinez-Jerez (2009). Relative performance compensation, contests, and dynamic incentives. *Management Science 55*(8), 1306–1320.

Chapman, J., M. Dean, P. Ortoleva, E. Snowberg, and C. Camerer (2017). Willingness to pay and willingness to accept are probably less correlated than you think. Working Paper 23954, National Bureau of Economic Research.

Chen, D. L., M. Schonger, and C. Wickens (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Working Paper 24678, National Bureau of Economic Research.

Czibor, E., D. Hsu, D. Jimenez-Gomez, S. Neckermann, and B. Subasi (2022). Loss-framed incentives and employee (mis-) behavior. *Management Science 68*(10), 7518–7537.

Davis, J. M. and S. B. Heller (2020, 10). Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs. *The Review of Economics and Statistics 102*(4), 664–677.

De Quidt, J. (2018). Your loss is my gain: a recruitment experiment with framed incentives. *Journal of the European Economic Association 16*(2), 522–559.

De Quidt, J., F. Fallucchi, F. Kölle, D. Nosenzo, and S. Quercia (2017). Bonus versus penalty: How robust are the effects of contract framing? *Journal of the Economic Science Association 3*(2), 174–182.

Delfgaauw, J., R. Dur, J. Sol, and W. Verbeke (2013). Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics 31*(2), 305–326.

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

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

Dohmen, T. and A. Falk (2011, April). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review 101*(2), 556–90.

Donato, K., G. Miller, M. Mohanan, Y. Truskinovsky, and M. Vera-Hernández (2017). Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in india. *American Economic Review 107*(5), 506–10.

Drouvelis, M. and M. L. Rigdon (2022). Gender differences in competitiveness: The role of social incentives. Working paper no. 9518, CESifo.

Dubé, J.-P. and S. Misra (2023). Personalized pricing and consumer welfare. *Journal of Political Economy 131*(1), 131–189.

Englmaier, F. and S. Leider (2020). Managerial payoff and gift-exchange in the field. *Review of Industrial Organization 56*(2), 259–280.

Eyring, H. and V. G. Narayanan (2018). Performance effects of setting a high reference point for peer-performance comparison. *Journal of Accounting Research 56*(2), 581–615.

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

Falk, A., A. Becker, T. J. Dohmen, D. Huffman, and U. Sunde (2022). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950.

Fallucchi, F., D. Nosenzo, and E. Reuben (2020). Measuring preferences for competition with experimentally-validated survey questions. *Journal of Economic Behavior & Organization 178*, 402–423.

Farrell, A. M., J. H. Grenier, and J. Leiby (2017). Scoundrels or stars? theory and evidence on the quality of workers in online labor markets. *The Accounting Review 92*(1), 93–114.

Farrell, M. H., T. Liang, and S. Misra (2021a). Deep learning for individual heterogeneity: An automatic inference framework. Working paper, arXiv:2010.14694.

Farrell, M. H., T. Liang, and S. Misra (2021b). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Ferraro, P. J. and J. D. Tracy (2022). A reassessment of the potential for loss-framed incentive contracts to increase productivity: A meta-analysis and a real-effort experiment. *Experimental Economics 25*, 1441–1466.

Foster, J. C., J. M. G. Taylor, and S. J. Ruberg (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine 30*(24), 2867–2880.

Friebel, G., M. Heinz, M. Krueger, and N. Zubanov (2017, August). Team incentives and performance: Evidence from a retail chain. *American Economic Review 107*(8), 2168–2203.

Fryer, R. G., S. D. Levitt, J. List, and S. Sadoff (2022). Enhancing the efficacy of teacher incentives through framing: A field experiment. *American Economic Journal: Economic Policy 14*(4), 269–299.

Gächter, S., E. J. Johnson, and A. Herrmann (2022). Individual-level loss aversion in riskless and risky choices. *Theory and Decision 92*, 599–624.

Gibbons, F. X. and B. P. Buunk (1999). Individual differences in social comparison: Development of a scale of social comparison orientation. *Journal of Personality and Social Psychology 76*(1), 129.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics 118*(3), 1049–1074.

Godinho de Matos, M., P. Ferreira, and M. D. Smith (2018). The effect of subscription video-on-demand on piracy: Evidence from a household-level randomized experiment. *Management Science 64*(12), 5610–5630.

Gosnell, G. K., J. A. List, and R. D. Metcalfe (2020). The impact of management practices on employee productivity: A field experiment with airline captains. *Journal of Political Economy 128*(4), 1195–1233.

Grolleau, G., M. G. Kocher, and A. Sutan (2016). Cheating and loss aversion: Do people cheat more to avoid a loss? *Management Science 62*(12), 3428–3438.

Grund, C. and D. Sliwka (2005). Envy and compassion in tournaments. *Journal of Economics & Management Strategy 14*(1), 187–207.

Hannan, R. L., V. B. Hoffman, and D. V. Moser (2005). Bonus versus penalty: Does contract frame affect employee effort? In A. Rapoport and R. Zwick (Eds.), *Experimental Business Research*, pp. 151–169. Boston, MA: Springer US.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Volume 2. Springer.

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*(5), 1683–1701.

Hitsch, G. J. and S. Misra (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. Working paper, Available at SSRN 3111957.

Horton, J. J., D. G. Rand, and R. J. Zeckhauser (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics 14*(3), 399–425.

Hossain, T. and J. A. List (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science 58*(12), 2151–2167.

Imai, K. and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics 7*(1), 443–470.

Imas, A. (2014). Working for the "warm glow": On the benefits and limits of prosocial incentives. *Journal of Public Economics 114*, 14–18.

Imas, A., S. Sadoff, and A. Samek (2017). Do people anticipate loss aversion? *Management Science 63*(5), 1271–1284.

John, O. P., E. M. Donahue, and R. L. Kentle (1991). Big five inventory–versions 4a and 54. Technical report, Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., L. P. Naumann, and C. J. Soto (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, and L. A. Pervin (Eds.), *Handbook of Personality: Theory and Research*, pp. 114–158. The Guilford Press.

Katz, L. F. and A. B. Krueger (2019). The rise and nature of alternative work arrangements in the united states, 1995–2015. *ILR Review 72*(2), 382–416.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics 133*(1), 237–293.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review 105*(5), 491–95.

Larkin, I. and S. Leider (2012, May). Incentive schemes, sorting, and behavioral biases of employees: Experimental evidence. *American Economic Journal: Microeconomics 4*(2), 184–214.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review 90*(5), 1346–1361.

Lazear, E. P. (2018). Compensation and incentives in the workplace. *Journal of Economic Perspectives 32*(3), 195–214.

Levitt, S. D., J. A. List, S. Neckermann, and S. Sadoff (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy 8*(4), 183–219.

Lourenço, S. M. (2020). Do self-reported motivators really motivate higher performance? *Management Accounting Research 47*, 100676.

Manthei, K., D. Sliwka, and T. Vogelsang (2021). Performance pay and prior learning—evidence from a retail chain. *Management Science 67*(11), 6998–7022.

Manthei, K., D. Sliwka, and T. Vogelsang (2023). Information provision, incentives, and attention: A field experiment on facilitating and influencing managers' decisions. *The Accounting Review 98*(5), 1–25.

Moreno-Torres, J. G., T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera (2012). A unifying view on dataset shift in classification. *Pattern recognition 45*(1), 521–530.

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

Ovadia, Y., E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems 32*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Quinonero-Candela, J., M. Sugiyama, A. Schwaighofer, and N. D. Lawrence (2008). *Dataset shift in machine learning*. Mit Press.

Rambachan, A., J. Kleinberg, J. Ludwig, and S. Mullainathan (2020, May). An economic perspective on algorithmic fairness. *AEA Papers and Proceedings 110*, 91–95.

Rammstedt, B. and O. P. John (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality 41*(1), 203–212.

Rau, H. A. (2014). The disposition effect and loss aversion: Do gender differences matter? *Economics Letters 123*(1), 33–36.

Skaperdas, S. and L. Gan (1995). Risk aversion in contests. *The Economic Journal 105*(431), 951–962.

Snowberg, E. and L. Yariv (2021). Testing the waters: Behavior across participant pools. *American Economic Review 111*(2), 687–719.

Sprinkle, G. B. and M. G. Williamson (2006). Experimental research in managerial accounting. *Handbooks of Management Accounting Research 1*, 415–444.

Stewart, N., C. Ungemach, A. J. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, and J. Chandler (2015). The average laboratory samples a population of 7,300 amazon mechanical turk workers. *Judgment and Decision Making 10*(5), 479–491.

Tonin, M. and M. Vlassopoulos (2010). Disentangling the sources of pro-socially motivated effort: A field experiment. *Journal of Public Economics 94*(11-12), 1086–1092.

Tonin, M. and M. Vlassopoulos (2015). Corporate philanthropy and productivity: Evidence from an online real effort experiment. *Management Science 61*(8), 1795–1811.

Van der Stede, W. A., A. Wu, and S. Y.-C. Wu (2020). An empirical analysis of employee responses to bonuses and penalties. *The Accounting Review 95*(6), 395–412.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Youyou, W., M. Kosinski, and D. Stillwell (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences 112*(4), 1036–1040.

A Online Appendix

A.1 Tables

Table A1: Summary Statistics

	Experiment 1		*Experiment 2*	
	Mean	S.D.	Mean	S.D.
Performance	1845.374	735.239	1962.573	723.225
Ability Proxy	39.946	23.247	43.042	21.993
Age	39.264	11.960	38.716	11.925
Female	0.464	0.499	0.493	0.500
Non-Binary	0.004	0.067	0.005	0.073
Some College	0.144	0.351	0.160	0.367
Bachelor's Degree or more	0.763	0.425	0.733	0.442
Observations	6065		6378	

Note: In this table, we report the summary statistics of Experiment 1 and Experiment 2. The ability proxy is measured as 'a/b'-presses participants reach in a 30 second test phase before they get their treatment description.

Table A2: Results of Experiment 1

	$log(Performance)_i$			
	(1)	(2)	(3)	(4)
PfP_i	0.375***	-0.029	-0.019	0.045
	(0.042)	(0.034)	(0.025)	(0.037)
$Bonus\ Gain_i$	0.359***	-0.044	-0.034	0.029
	(0.050)	(0.044)	(0.046)	(0.051)
$Gift\ and\ Goal_i$	0.210***	-0.194***	-0.184***	-0.120**
	(0.052)	(0.048)	(0.044)	(0.048)
$Bonus\ Loss_i$	0.403***		0.010	0.073*
	(0.047)		(0.034)	(0.039)
$RTR\ Feedback_i$	0.394***	-0.010		0.064*
	(0.042)	(0.034)		(0.036)
$Social\ PfP_i$	0.330***	-0.073*	-0.064*	
	(0.051)	(0.039)	(0.036)	
$Control_i$		-0.403***	-0.394***	-0.330***
		(0.047)	(0.042)	(0.051)
Observations	6,065	6,065	6,065	6,065
Adjusted R-squared	0.125	0.125	0.125	0.125

Note: In this table, we report the results of regressions of log(Performance) on treatment dummies for all but one treatment in Experiment 1. In column (1), we use the control group as reference group, thus reporting the treatment effects for the different incentive schemes in comparison to the control group. In column (2) to (4), we use the *Bonus Loss*, the *Real-time Rank Feedback* and the *Social PfP* treatment as reference group, respectively. We include batch fixed effects as well as an ability proxy as control.The ability proxy is measured as 'a/b'-presses participants reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Algorithm Comparison

	Residualized Performance			
	Bonus Loss	RTR Feedback	Social PfP	Overall
	(1)	(2)	(3)	(4)
Indirect Random Forest	126.9	133.3	154.0	132.7
(Share of Obs.)	(46.8)	(41.6)	(11.6)	(100.0)
Causal Forest	130.1	120.1	115.5	125.7
(Share of Obs.)	(55.6)	(39.0)	(5.4)	(100.0)
Doubly Robust	130.3	134.2	102.1	129.8
(Share of Obs.)	(52.9)	(39.9)	(7.2)	(100.0)
Causal Net	121.7	133.9	58.6	120.0
(Share of Obs.)	(56.8)	(33.9)	(9.2)	(100.0)
All	112.6	87.6	30.5	

Note: In this table, we report the average residualized performance of workers in the *Bonus Loss* treatment (column (1)), in the *RTR Feedback* treatment (column (2)), in the *Social PfP* treatment (column (3)) or in any of these treatments (column (4)) who were randomly allocated to their predictably best incentive scheme in Experiment 1. We residualized performance on the ability proxy. The ability proxy is measured as 'a/b'-presses participants reach in a 30 second test phase before they get their treatment description. We compute the average over the residualized performance of 50 runs of a 3-fold cross-validation where we predict the best incentive scheme out-of-sample. We report the results for four different algorithms (Indirect Random Forest, Causal Forest, Doubly Robust and Causal Net). We report the percent of observations coming from the different treatments when computing the average overall in parenthesis. We also report the average residualized performance of all workers in the the treatments independent of their predictably best treatment ("All").

Table A4: Feature Importances

Features	Bonus Loss	RTR Feedback	Social PfP
Age	.159	.189	.232
Gender	.014	.039	.05
Education	.011	.003	.006
Conscientiousness - Item 1	.024	.007	.016
Conscientiousness - Item 2 (rev)	.053	.017	.052
Conscientiousness - Item 3	.011	.006	.009
Conscientiousness - Item 4	.014	.003	.024
Agreeableness - Item 1	.019	.015	.006
Agreeableness - Item 2 (rev)	.021	.016	.056
Emotional Stability - Item 1	.015	.009	.011
Emotional Stability - Item 2 (rev)	.02	.007	.011
Openness - Item 1 (rev)	.015	.025	.025
Openness - Item 2	.024	.003	.034
Extraversion - Item 1 (rev)	.014	.008	.011
Extraversion - Item 2	.027	.063	.011
Altruism - Item 1	.238	.39	.117
Altruism - Item 2	.036	.008	.023
Risk Aversion - Item 1 (rev)	.04	.081	.045
Positive Reciprocity - Item 1	.037	.017	.018
Positive Reciprocity - Item 2	.019	.007	.121
Social Comparison - Item 1	.014	.005	.011
Social Comparison - Item 2 (rev)	.028	.015	.022
Social Comparison - Item 3	.014	.018	.021
Competitiveness - Item 1	.033	.017	.02
Competitiveness - Item 2	.042	.016	.026
Competitiveness - Item 3	.043	.011	.014
Competitiveness - Item 4	.015	.005	.007

Note: In this table, we report the relative feature importance for the second stage models of the indirect random forest approach predicting the CATEs for the three incentive schemes. We compute the feature importance as Gini importance, i.e. using the loss reduction at each internal node of each tree. See, for example, chapter 10 of Hastie et al. (2009) for details. Using permutation-based importance (Breiman, 2001), i.e. randomly reshuffling each feature and computing the resulting loss increase, led to qualitatively same results.

Table A5: Treatment Share Balance - Retaker

	$Retaker_i$	
	(1)	(2)
$Algorithm_i$	-0.026	-0.024
	(0.028)	(0.019)
$Best\ ATE_i$	-0.014	-0.019
	(0.030)	(0.019)
Controls	No	Yes
Observations	6,378	6,377
Adjusted R-squared	-0.000	0.571

Note: In this table, we report the results of regressions of a Retaker dummy on treatment dummys for the *Best ATE* treatment as well as the *Algorithm* treatment. In columns (2), include batch fixed effects as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses participants reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Summary Statistics - Unit Cost

	Best ATE	Algorithm	Random Assignment
	(1)	(2)	(3)
Bonus Loss			
Cost per Worker (USD)	0.71	0.64	0.68
Cost per Unit (USD)	0.00036	0.00034	0.00034
RTR Feedback			
Cost per Worker (USD)	-	1.28	1.15
Cost per Unit (USD)	-	0.00061	0.00059
Social PfP			
Cost per Worker (USD)	-	0.95	0.92
Cost per Unit (USD)	-	0.00049	0.00049
Overall			
Cost per Worker (USD)	0.71	0.98	0.92
Cost per Unit (USD)	0.00036	0.00050	0.00048

Note: In this table, we report the mean costs per worker as well as the mean costs per unit for the different assignments as well as incentive schemes.

Table A7: Robustness Check: Effect on Performance (Retaker Control)

	$log(Performance)_i$		
	All	All	Retakers
	(1)	(2)	(3)
$Algorithm_i$	0.257***	0.043**	0.097***
	(0.056)	(0.017)	(0.028)
$Best\ ATE_i$	0.215***		
	(0.057)		
$Retaker_i$	-0.077	-0.080	
	(0.054)	(0.051)	
$Retaker_i \times Exp1\ PfP_i$	0.003	-0.006	-0.010
	(0.074)	(0.071)	(0.076)
$Retaker_i \times Exp1\ Bonus\ Gain_i$	0.049	0.032	0.039
	(0.056)	(0.055)	(0.056)
$Retaker_i \times Exp1\ Gift\ and\ Goal_i$	0.051	0.027	0.027
	(0.063)	(0.064)	(0.068)
$Retaker_i \times Exp1\ Bonus\ Loss_i$	0.041	0.022	0.019
	(0.090)	(0.088)	(0.092)
$Retaker_i \times Exp1\ RTR\ Feedback_i$	-0.041	-0.041	-0.050
	(0.089)	(0.089)	(0.093)
$Retaker_i \times Exp1\ Social\ PfP_i$	-0.082	-0.097	-0.105
	(0.072)	(0.070)	(0.074)
p-value $Best\ ATE_i$=$Algorithm_i$	0.014		
Reference Group	Control	Best ATE	Best ATE
Observations	6,377	6,147	2,015
Adjusted R-squared	0.112	0.111	0.132

Note: In this table, we report the results of regressions of log(Performance) on treatment dummys for the *Best ATE* treatment as well as the *Algorithm* treatment. In columns (2) and (3), we exclude the control group so that *Best ATE* is the reference group for the *Algorithm* dummy. In column (3), we further restrict the sample to Retakers. We include batch fixed effects as well as an ability proxy, a Retaker dummy and dummies for the Retakers' treatments in Experiment 1 as controls. The ability proxy is measured as 'a/b'-presses participants reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Relationship between Experiment 1 Treatment and Retaking

	$Retaker_i$	
	Logit	Fixed Effects Regression
	(1)	(2)
PfP_i	0.000	-0.001
	(0.114)	(0.026)
$Bonus\ Gain_i$	0.080	0.018
	(0.123)	(0.028)
$Gift\ and\ Goal_i$	0.107	0.023
	(0.122)	(0.028)
$Bonus\ Loss_i$	0.002	0.001
	(0.100)	(0.022)
$RTR\ Feedback_i$	0.009	0.002
	(0.100)	(0.022)
$Social\ PfP_i$	0.039	0.008
	(0.123)	(0.027)
Observations	6,065	6,065
Pseudo R-squared	0.000	
Adjusted R-squared		0.012

Note: In this table, we report the results of logistic regressions (columns (1)) and fixed effect regressions (columns (2)) of a Retaker dummy on dummies for the treatment in Experiment 1. The control treatment serves as reference group. For the fixed effects regression we use a linear regression with batch fixed effects. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Robustness Check: Effect on Performance

	$Performance_i$			
	All	All	New Hires	Retakers
	(1)	(2)	(3)	(4)
$Algorithm_i$	343.073***	31.325**	9.833	70.716**
	(53.865)	(13.799)	(17.420)	(30.844)
$Best\ ATE_i$	311.671***			
	(53.617)			
p-value $Best\ ATE_i$=$Algorithm_i$	0.027			
Reference Group	Control	Best ATE	Best ATE	Best ATE
Observations	6,377	6,147	4,131	2,015
Adjusted R-squared	0.178	0.173	0.172	0.182

Note: In this table, we report the results of regressions of Performance, i.e. the number of achieved points, on treatment dummys for the *Best ATE* treatment as well as the *Algorithm* treatment. In columns (2)-(4), we exclude the control group so that *Best ATE* is the reference group for the *Algorithm* dummy. We include batch fixed effects as well as an ability proxy as control. The ability proxy is measured as 'a/b'-presses participants reach in a 30 second test phase before they get their treatment description. In columns (3) and (4), we restrict the sample to New Hires and Retakers, respectively. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A10: Subgroup Analysis Separately for Retakers and New Hires

	$log(Performance)_i$		
	Predicted Bonus Loss	Predicted RTR Feedback	Predicted Social PfP
	(1)	(2)	(3)
Panel A: New Hires			
$Algorithm_i$	-0.009	0.012	0.030
	(0.057)	(0.030)	(0.069)
Reference Group	Best ATE	Best ATE	Best ATE
Observations	1,612	1,950	564
Adjusted R-squared	0.084	0.108	0.151
Panel B: Retakers			
$Algorithm_i$	0.037	0.175***	-0.034
	(0.062)	(0.042)	(0.138)
Reference Group	Best ATE	Best ATE	Best ATE
Observations	818	955	232
Adjusted R-squared	0.137	0.101	0.161

Note: In this table, we report the results of regressions of log(Performance) on an *Algorithm* treatment dummy in sub-samples split by the predicted best treatment. We exclude the control group so that *Best ATE* is the reference group for the *Algorithm* dummy. Panel A shows the results for the sample of New Hires and Panel B for the sample of Retakers, respectively. Column (1) presents the results for the sub-sample of all participants (regardless of their actual assignment) for which the *Bonus Loss* was predicted to be the best incentive scheme based on their individual characteristics. Column (2) and (3) present the results for the sub-sample of all workers (regardless of their actual assignment) for which the *Real-time Rank Feedback* and *Social PfP* was predicted to be the best incentive scheme based on their individual characteristics, respectively. We further include batch fixed effects and an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Group Characteristics (Logit) - New Hires

	Predicted Bonus Loss$_i$ (1)	*Predicted RTR Feedback*$_i$ (2)	*Predicted Social PfP*$_i$ (3)
Age$_i$	0.083***	-0.021***	-0.248***
	(0.005)	(0.004)	(0.016)
Female$_i$	1.326***	-0.929***	-1.498***
	(0.132)	(0.109)	(0.221)
Some College$_i$	-0.184	0.226	-0.523
	(0.135)	(0.144)	(0.364)
Bachelor's Degree or more$_i$	-0.072	-0.140	0.598**
	(0.138)	(0.154)	(0.251)
Ability Proxy$_i$	-0.122***	0.243***	-0.137***
	(0.038)	(0.048)	(0.048)
Conscientiousness$_i$	0.387***	-0.229***	-0.623***
	(0.072)	(0.064)	(0.089)
Openness$_i$	-0.241***	0.164***	0.119
	(0.049)	(0.051)	(0.096)
Emotional Stability$_i$	0.093	-0.090	-0.136
	(0.076)	(0.088)	(0.101)
Agreeableness$_i$	-0.043	0.064	0.169**
	(0.068)	(0.074)	(0.082)
Extraversion$_i$	0.384***	-0.551***	0.566***
	(0.054)	(0.069)	(0.108)
Altruism$_i$	0.658***	-1.972***	2.099***
	(0.081)	(0.100)	(0.145)
Positive Reciprocity$_i$	-0.563***	0.452***	0.314***
	(0.063)	(0.069)	(0.105)
Competitiveness$_i$	1.066***	-1.103***	-0.020
	(0.075)	(0.061)	(0.146)
Social Comparison$_i$	0.231***	-0.096	-0.404***
	(0.070)	(0.071)	(0.112)
Risk Aversion$_i$	-0.706***	0.629***	0.041
	(0.056)	(0.065)	(0.090)
Observations	4,282	4,282	4,282
Pseudo R-squared	0.320	0.447	0.454

Note: In this table, we report the results of a logistic regression of a dummy of having *Bonus Loss* (column (1)), *RTR Feedback* (column (2)), or *Social PfP* (column (3)) as predicted best incentive scheme for the New Hires sample on the features the algorithm uses for assignment. With the exception of age (continuous), female (binary), some college (binary) and bachelor's degree or more (binary) all variables are standardized. For all characteristics for which we used more than one item as a feature, we built a summative scale (i.e. for the big-5, altruism, positive reciprocity, competitiveness and social comparison). Standard errors are clustered at the batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Gender Pay Disparity

	$Payout_i$			
	All	Bonus Loss	RTR Feedback	Social PfP
	(1)	(2)	(3)	(4)
$Female_i$	-0.003	-0.002	0.000	-0.001
	(0.015)	(0.015)	(0.015)	(0.015)
$Algorithm_i$	0.326***	-0.055***	0.594***	0.234***
	(0.026)	(0.019)	(0.021)	(0.027)
$Female_i \times Algorithm_i$	-0.091***	0.037	-0.127***	0.057*
	(0.025)	(0.024)	(0.031)	(0.032)
Observations	6,115	4,269	4,531	3,459
Adjusted R-squared	0.254	0.162	0.371	0.188

Note: In this table, we report the results of regressions of Payout, i.e. the bonus workers received for the working task, on a dummy for being female, a dummy for being in the *Algorithm* treatment and the interaction between both. The reference group is the *Best ATE*, i.e. the *Bonus Loss* treatment. We exclude the control group as well as workers identifying as non-binary in this analysis. Column (1) presents results including all workers in the *Algorithm* and *Best ATE* treatments, In the remaining columns, we include only those workers from the *Algorithm* treatment that are assigned to *Bonus Loss* (column (2)), *RTR Feedback* (column (3)) and *Social PfP* (column (4)), respectively. We include batch fixed effects as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: Age Group Pay Disparity

	$Payout_i$			
	All	Bonus Loss	RTR Feedback	Social PfP
	(1)	(2)	(3)	(4)
$Algorithm_i$	0.364***	-0.062	0.589***	0.250***
	(0.029)	(0.044)	(0.032)	(0.030)
Between 31 and 40$_i$	0.053***	0.050***	0.052***	0.048***
	(0.017)	(0.016)	(0.016)	(0.016)
Between 41 and 50$_i$	0.038	0.039	0.041	0.039
	(0.026)	(0.025)	(0.025)	(0.026)
Older Than 50$_i$	-0.045*	-0.043*	-0.034	-0.044*
	(0.025)	(0.024)	(0.025)	(0.025)
Between 31 and 40$_i \times Algorithm_i$	-0.091***	0.044	-0.061	0.014
	(0.032)	(0.052)	(0.041)	(0.035)
Between 41 and 50$_i \times Algorithm_i$	-0.188***	0.003	-0.107**	0.136***
	(0.035)	(0.062)	(0.042)	(0.031)
Older Than 50$_i \times Algorithm_i$	-0.075*	0.053	-0.059	0.174***
	(0.041)	(0.060)	(0.049)	(0.034)
Observations	6,147	4,288	4,557	3,478
Adjusted R-squared	0.258	0.166	0.371	0.192

Note: In this table, we report the results of regressions of Payout, i.e. the bonus workers received for the working task, on a dummy for being in the *Algorithm* treatment, dummies for being in a certain age group and the interaction between both. The reference group is the age group below 31 in the *Best ATE*, i.e. the *Bonus Loss* treatment. Column (1) presents results including all workers in the *Algorithm* and *Best ATE* treatments, In the remaining columns, we include only those workers from the *Algorithm* treatment that are assigned to *Bonus Loss* (column (2)), *RTR Feedback* (column (3)) and *Social PfP* (column (4)), respectively. We include batch fixed effects as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: Sample Differences between Exp1, Retakers and New Hires

	Exp 1		*Retakers*		*New Hires*		Exp1-NH	Ret-NH
	Mean	S.D.	Mean	S.D.	Mean	S.D.	p-value	p-value
Female	0.464	0.499	0.451	0.498	0.513	0.500	0.000	0.000
Some College	0.144	0.351	0.141	0.348	0.170	0.375	0.000	0.004
Bachelor's Degree or more	0.763	0.425	0.771	0.420	0.715	0.452	0.000	0.000
Age	39.264	11.960	40.711	12.250	37.740	11.640	0.000	0.000
Ability Proxy	39.946	23.247	46.054	20.435	41.568	22.573	0.000	0.000
Conscientiousness	0.012	0.702	0.094	0.739	-0.063	0.700	0.000	0.000
Agreeableness	0.010	0.719	0.043	0.780	-0.035	0.706	0.002	0.000
Openness	-0.008	0.736	0.078	0.768	-0.026	0.731	0.219	0.000
Emotional Stability	0.014	0.798	0.103	0.845	-0.070	0.779	0.000	0.000
Extraversion	0.004	0.799	0.008	0.856	-0.010	0.765	0.370	0.428
Altuism	0.019	0.747	-0.019	0.773	-0.017	0.768	0.019	0.906
Risk Aversion	-0.020	1.008	0.109	1.036	-0.025	0.966	0.812	0.000
Positive Reciprocity	-0.003	0.787	-0.004	0.804	0.006	0.760	0.594	0.653
Competitiveness	0.021	0.806	-0.084	0.852	0.011	0.783	0.497	0.000
Social Comparison	-0.009	0.676	-0.079	0.747	0.052	0.628	0.000	0.000
Observations	6065		2096		4282		10347	6378

Note: In this table, we report the summary statistics of Experiment 1 as well as Retakers and New Hires in Experiment 2. Moreover, we report the p-values of a t-test for the continuous variables age, ability proxy and z-scored personality traits as well as social and economic preferences, testing the null hypothesis whether the samples are the same. The ability proxy is measured as 'a/b'-presses participants reach in a 30 second test phase before they get their treatment description. For the binary variables, we present the p-values of a test of proportions, testing the null hypothesis whether the samples are the same.

Table A15: Robustness: Mechanisms - New Hires

	$log(Performance)_i$			
	Training Similarity (1)	Similarity & Fut. Int. (2)	Similarity & Consistency (3)	All (4)
$Algorithm_i$	0.052	0.110***	0.122***	0.152***
	(0.044)	(0.041)	(0.037)	(0.038)
$\times$ *Training Similarity Q2*$_i$	-0.041	-0.030	-0.048	-0.039
	(0.052)	(0.053)	(0.053)	(0.053)
$\times$ *Training Similarity Q3*$_i$	-0.047	-0.027	-0.051	-0.035
	(0.063)	(0.067)	(0.063)	(0.066)
$\times$ *Training Similarity Q4*$_i$	-0.056	-0.014	-0.053	-0.022
	(0.051)	(0.053)	(0.053)	(0.052)
$\times$ *Future Interaction Q2*$_i$		-0.061		-0.048
		(0.051)		(0.049)
$\times$ *Future Interaction Q3*$_i$		-0.086*		-0.066
		(0.048)		(0.045)
$\times$ *Future Interaction Q4*$_i$		-0.165***		-0.122**
		(0.059)		(0.054)
$\times$ *Consistency Q2*$_i$			0.012	0.019
			(0.031)	(0.032)
$\times$ *Consistency Q3*$_i$			-0.123***	-0.108**
			(0.044)	(0.043)
$\times$ *Consistency Q4*$_i$			-0.163**	-0.130**
			(0.062)	(0.061)
Reference Group	Best ATE	Best ATE	Best ATE	Best ATE
Observations	4,131	4,131	4,131	4,131
Adjusted R-squared	0.098	0.100	0.101	0.101

Note: In this table, we report the results of regressions of log(Performance) on a dummy for being in the *Algorithm* treatment as well as interactions between *Algorithm* and being in the second, third or lowest quartile of the New Hire sample with regards to their predicted similarity with the training sample. We exclude the control group so that the *Best ATE* treatment group is the reference group for the *Algorithm* dummy. In column (2), we include additionally the interactions between *Algorithm* and being in the second, third or lowest quartile of the New Hire sample with regards to their predicted propensity for future interaction. In column (3), we include additionally interactions between *Algorithm* and being in the second, third or lowest quartile regarding consistency. In column (4), we additionally include predicted propensity for future interactions as well as consistency quartiles. We include batch fixed effects as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A16: Consistency Comparison between Retakers and Other Workers

	Z-scored Consistency$_i$	
	Experiment 2	Experiment 1
	(1)	(2)
Retaker (Exp2)$_i$	0.290***	0.241***
	(0.039)	(0.027)
Observations	6,377	6,065
Adjusted R-squared	0.121	0.137

Note: In this table, we report the results of regressions of consistency in survey answers on a dummy for being a Retaker in Experiment 2. The measure for the consistency of survey answers is the z-scored reversed mean absolute distance between mean answers to originally reversed-coded and normally coded items of the measured characteristics (after reversing the scales so that they are coded in the same direction). In column (1), we restrict the sample to experiment 2, i.e. the reference group for the Retakers (Experiment 2) are the New Hires. In column (2), we restrict the sample to Experiment 1, i.e. the reference group for the Retakers (Experiment 2) are the workers taking part in Experiment 1 only. We include batch fixed effects, as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A17: Algorithm Effect by Propensity for Future Interaction

	$log(Performance)_i$		
	Future Interaction Prop. ≥ 40%	Future Interaction Prop. ≥ 50%	Future Interaction Prop. ≥ 60%
	(1)	(2)	(3)
Panel A: New Hires			
$Algorithm_i$	0.026	0.048	0.117*
	(0.031)	(0.041)	(0.069)
Reference Group	Best ATE	Best ATE	Best ATE
Observations	1,671	1,052	434
Adjusted R-squared	0.096	0.086	0.078
Panel B: Retakers			
$Algorithm_i$	0.094***	0.094***	0.094***
	(0.028)	(0.028)	(0.028)
Reference Group	Best ATE	Best ATE	Best ATE
Observations	2,014	2,014	2,014
Adjusted R-squared	0.130	0.130	0.130

Note: In this table, we report the results of regressions of log(Performance) on a dummy for being in the *Algorithm* treatment. We exclude the control group so that the *Best ATE* treatment group is the reference group for the *Algorithm* dummy. Panel A shows results for the sample of New Hires and Panel B for the sample of Retakers. Column (1) presents the results for a subsample with a predicted propensity for future interaction of at least the fortieth percentile of the full sample (i.e. Retakers and New Hires together). Column (2) and column (3) show the results for the fiftieth and sixtieth percentiles, respectively. We predict propensity for future interaction using a simple random forest model trained on the Experiment 1 data including information who became a Retaker later on. We include batch fixed effects as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A18: Robustness: Consistency - Retaker

	$log(Performance)_i$		
	Consistency	Test-Retest Reliability	Both
	(1)	(2)	(3)
$Algorithm_i$	0.146***	0.194***	0.182***
	(0.046)	(0.041)	(0.059)
$\times$ $Consistency\ Q2_i$	-0.019		0.030
	(0.072)		(0.067)
$\times$ $Consistency\ Q3_i$	-0.034		0.038
	(0.072)		(0.057)
$\times$ $Consistency\ Q4_i$	-0.147**		-0.023
	(0.068)		(0.087)
$\times$ $Test\text{-}Retest\ Rel.\ Q2_i$		0.006	0.003
		(0.043)	(0.039)
$\times$ $Test\text{-}Retest\ Rel.\ Q3_i$		-0.183***	-0.180***
		(0.037)	(0.043)
$\times$ $Test\text{-}Retest\ Rel.\ Q4_i$		-0.217**	-0.211*
		(0.103)	(0.118)
Reference Group	Best ATE	Best ATE	Best ATE
Observations	2,015	2,015	2,015
Adjusted R-squared	0.133	0.136	0.135

Note: In this table, we report the results of regressions of log(Performance) on a dummy for being in the *Algorithm* treatment. In column (1), we include additionally interactions between *Algorithm* and being in the second, third or lowest quartile regarding consistency in the Retaker sample. The measure for the consistency of survey answers is the z-scored reversed mean absolute distance between mean answers to originally reversed-coded and normally coded items of the measured characteristics (after reversing the scales so that they are coded in the same direction). In column (2), we further include interactions between *Algorithm* and being in the second, third or lowest quartile regarding test-retest reliability in the Retaker sample. We measure test-retest reliability as Spearman's rank correlation between the survey answers in the first and in the second experiment. We exclude answers regarding demographics and one of the altruism items as the answer is a monetary value between 1 and 1000 while all other questions are answered on ordinal scales. In column (2), we further include consistency as well as test-retest reliability quartiles. We include batch fixed effects as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A19: Algorithm Effect by Consistency

	$log(Performance)_i$		
	Consistency ≥ 40%	Consistency ≥ 50%	Consistency ≥ 60%
	(1)	(2)	(3)
Panel A: New Hires			
$Algorithm_i$	0.021	0.048*	0.059**
	(0.022)	(0.025)	(0.029)
Reference Group	Best ATE	Best ATE	Best ATE
Observations	2,423	2,007	1,548
Adjusted R-squared	0.100	0.097	0.106
Panel B: Retakers			
$Algorithm_i$	0.096***	0.119***	0.102***
	(0.031)	(0.035)	(0.035)
Reference Group	Best ATE	Best ATE	Best ATE
Observations	1,322	1,126	928
Adjusted R-squared	0.096	0.096	0.092

Note: In this table, we report the results of regressions of log(Performance) on a dummy for being in the *Algorithm* treatment. We exclude the control group so that the *Best ATE* treatment group is the reference group for the *Algorithm* dummy. Panel A shows results for the sample of New Hires and Panel B for the sample of Retakers. Column (1) presents the results for a subsample with a consistency of at least the fortieth percentile of the full sample (i.e. Retakers and New Hires together). Column (2) and column (3) show the results for the fiftieth and sixtieth percentiles, respectively. The measure for the consistency of survey answers is the z-scored reversed mean absolute distance between mean answers to originally reversed-coded and normally coded items of the measured characteristics (after reversing the scales so that they are coded in the same direction). We include batch fixed effects as well as an ability proxy as controls. The ability proxy is measured as 'a/b'-presses workers reach in a 30 second test phase before they get their treatment description. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A20: Comparison of Effect in Experiment 1 and Experiment 2 (Retakers)

	$log(Performance)_i$	
	Experiment 1	Experiment 2
	(1)	(2)
$Algorithm_i$	0.113***	0.097***
	(0.037)	(0.028)
Reference Group	Best ATE	Best ATE
Observations	585	2,015
Adjusted R-squared	0.125	0.132

Note: In this table, we report the results of regressions of log(Performance) on a dummy for being by chance in the by the algorithm predicted best treatment (column (1), Experiment 1) or in the *Algorithm* treatment (column (2), Experiment 2). The sample is restricted to the Retakers, i.e. those participants who take part in experiment 1 and in experiment 2. The reference group is in both cases the *Best ATE*, i.e. the *Bonus Loss* treatment. All other treatments are excluded. Observations of Retakers who are by chance in *Bonus Loss* in Experiment 1 and for whom *Bonus Loss* is also the predicted best treatment are duplicated and used in both samples. Standard errors are clustered on batch level, and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.2 Figures

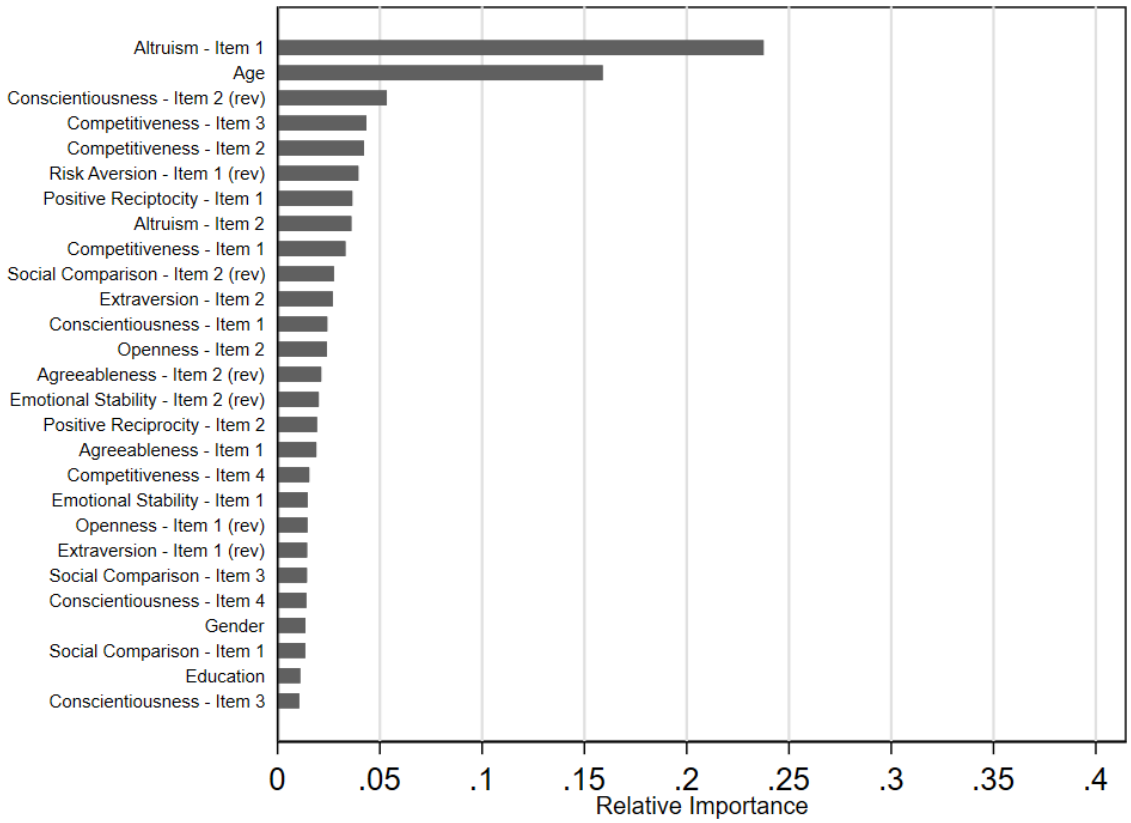


Figure A1: Feature Importances - Bonus Loss

Note: This figure shows the relative feature importance for the second stage model of the indirect random forest approach predicting the CATE for the *Bonus Loss* incentive scheme. We compute the feature importance as Gini importance, i.e. using the loss reduction at each internal node of each tree. See Table A4 in the Online Appendix for the exact values. For details, see, for example, chapter 10 of Hastie et al. (2009). Using permutation-based importance (Breiman, 2001), i.e. randomly reshuffling each feature and computing the resulting loss increase, led to qualitatively same results.

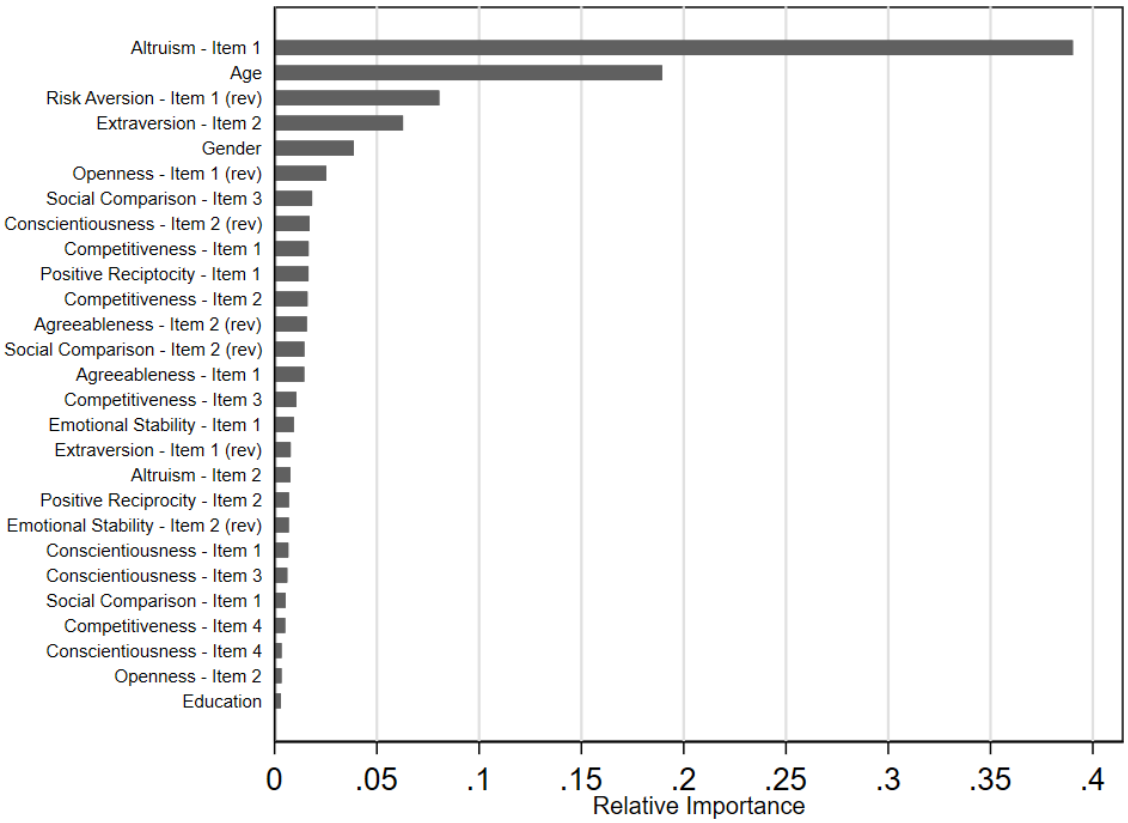


Figure A2: Feature Importances - RTR Feedback

Note: This figure shows the relative feature importance for the second stage model of the indirect random forest approach predicting the CATE for the *RTR Feedback* incentive scheme. We compute the feature importance as Gini importance, i.e. using the loss reduction at each internal node of each tree. See Table A4 in the Online Appendix for the exact values. For details, see, for example, chapter 10 of Hastie et al. (2009). Using permutation-based importance (Breiman, 2001), i.e. randomly reshuffling each feature and computing the resulting loss increase, led to qualitatively same results.

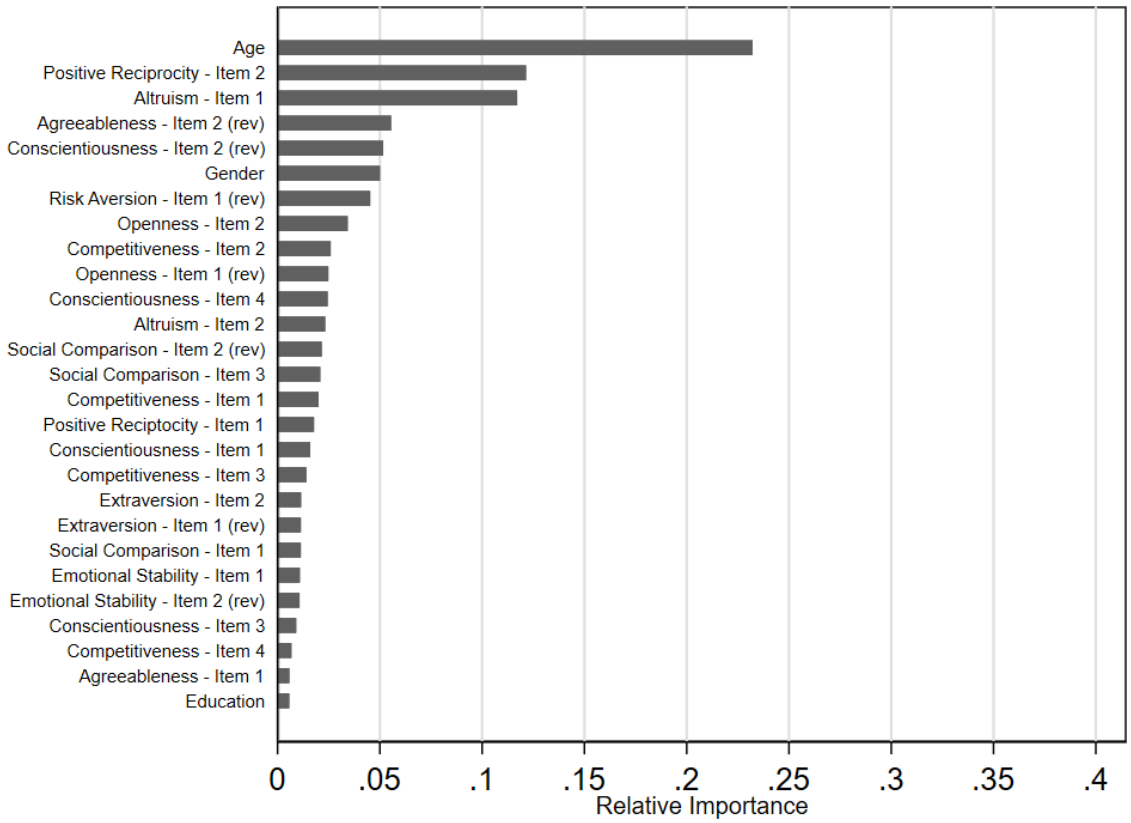
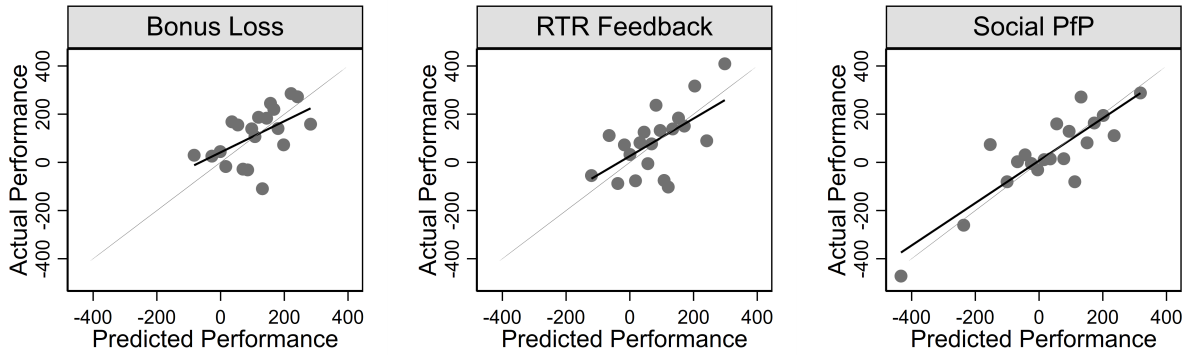
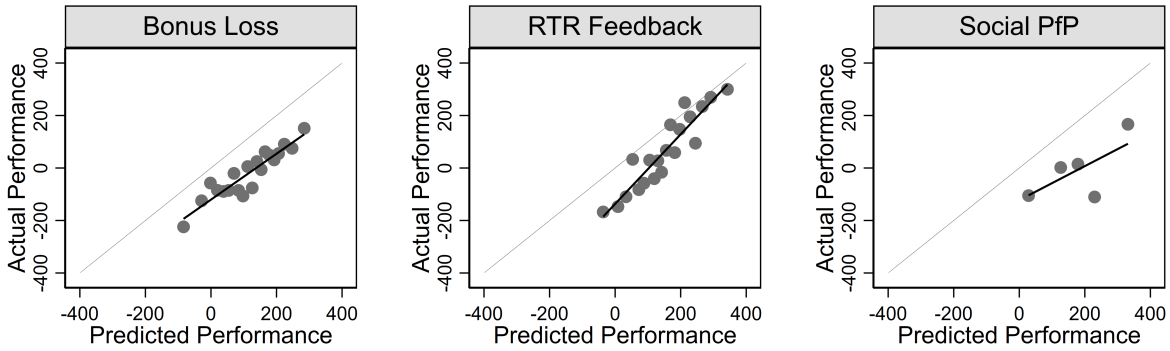


Figure A3: Feature Importances - Social PfP

Note: This figure shows the relative feature importance for the second stage model of the indirect random forest approach predicting the CATE for the *Social PfP* incentive scheme. We compute the feature importance as Gini importance, i.e. using the loss reduction at each internal node of each tree. See Table A4 in the Online Appendix for the exact values. For details, see, for example, chapter 10 of Hastie et al. (2009). Using permutation-based importance (Breiman, 2001), i.e. randomly reshuffling each feature and computing the resulting loss increase, led to qualitatively same results.



(a) Experiment 1



(b) Experiment 2

Figure A4: Predicted vs Actual Performance

Note: This figure shows binned scatterplots for the predicted vs actual performance for the *Bonus Loss*, *RTR Feedback* and *Social PfP* treatments in the first experiment (panel (a)) and the second experiment (panel (b)). We predict the performance out-of-sample using the first stage of our chosen indirect RF algorithm and 10-fold cross-validation. We also show the linear fit line of a regression of actual performance on predicted performance.

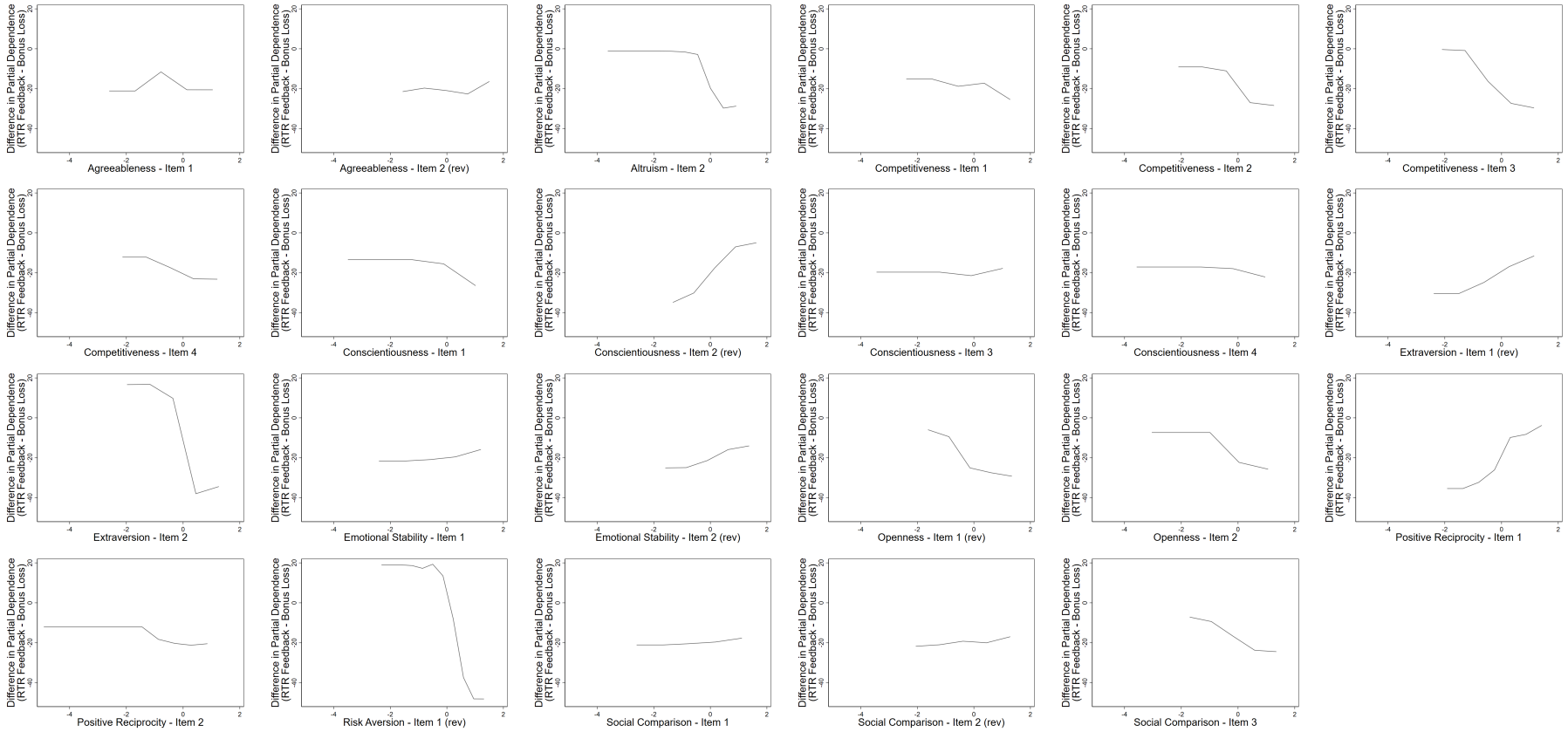


Figure A5: Partial Dependence Comparisons (RTR Feedback - Bonus Loss)

Note: This figure shows the difference in partial dependence between the *RTR Feedback* scheme and the *Bonus Loss* scheme (i.e. the incentive scheme with the highest point estimate in the first experiment) for all characteristics passed to the algorithm as features, with the exception of demographics and an item measuring altruism (see Figure 2 in Section 3 for the figures for age and the altruism item). All characteristics are z-scored.

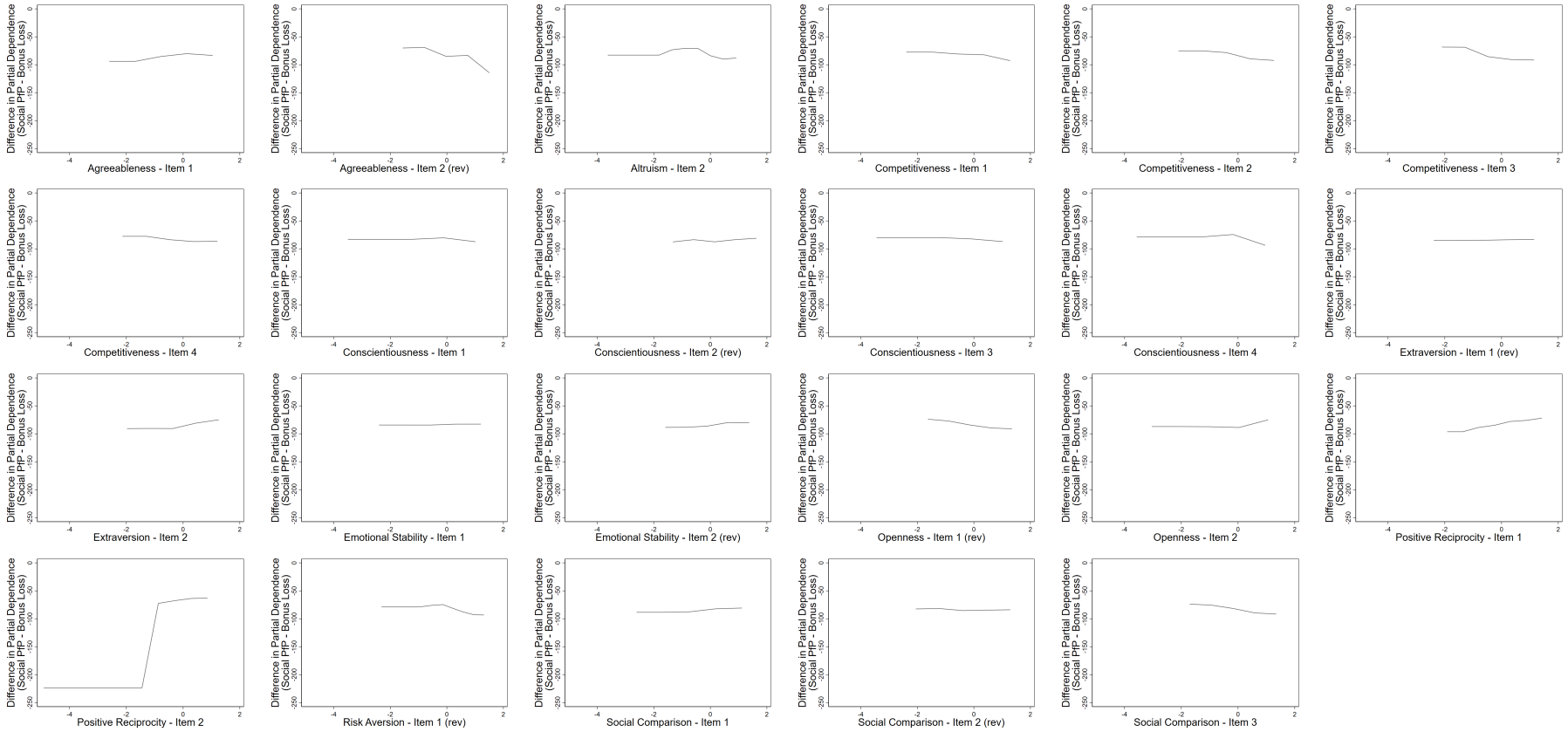


Figure A6: Partial Dependence Comparisons (Social PfP - Bonus Loss)

Note: This figure shows the difference in partial dependence between the *Social PfP* scheme and the *Bonus Loss* scheme (i.e. the incentive scheme with the highest point estimate in the first experiment) for all characteristics passed to the algorithm as features, with the exception of demographics and an item measuring altruism (see Figure 2 in Section 3 for the figures for age and the altruism item). All characteristics are z-scored.

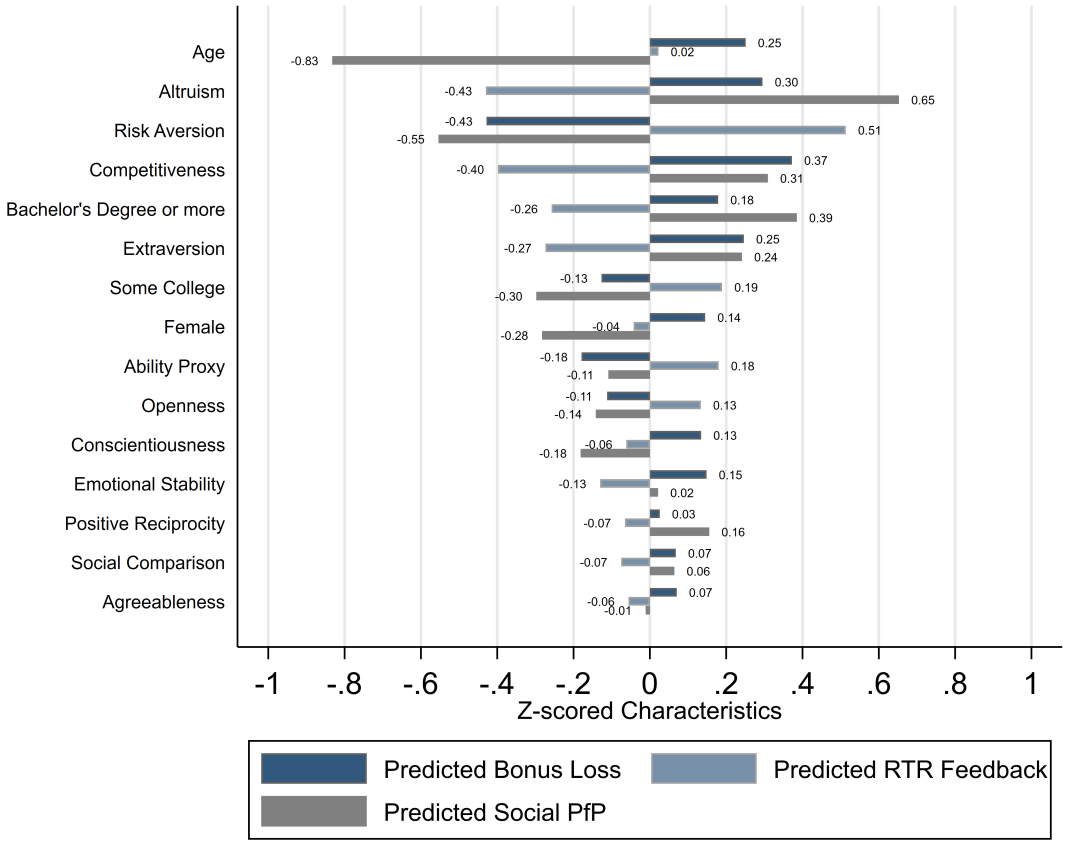


Figure A7: Group Characteristics

Note: This figure shows the averages of each characteristic in the three groups resulting from a split depending on the predictably best treatment in Experiment 2. All characteristics are z-scored.

A.3 Instructions

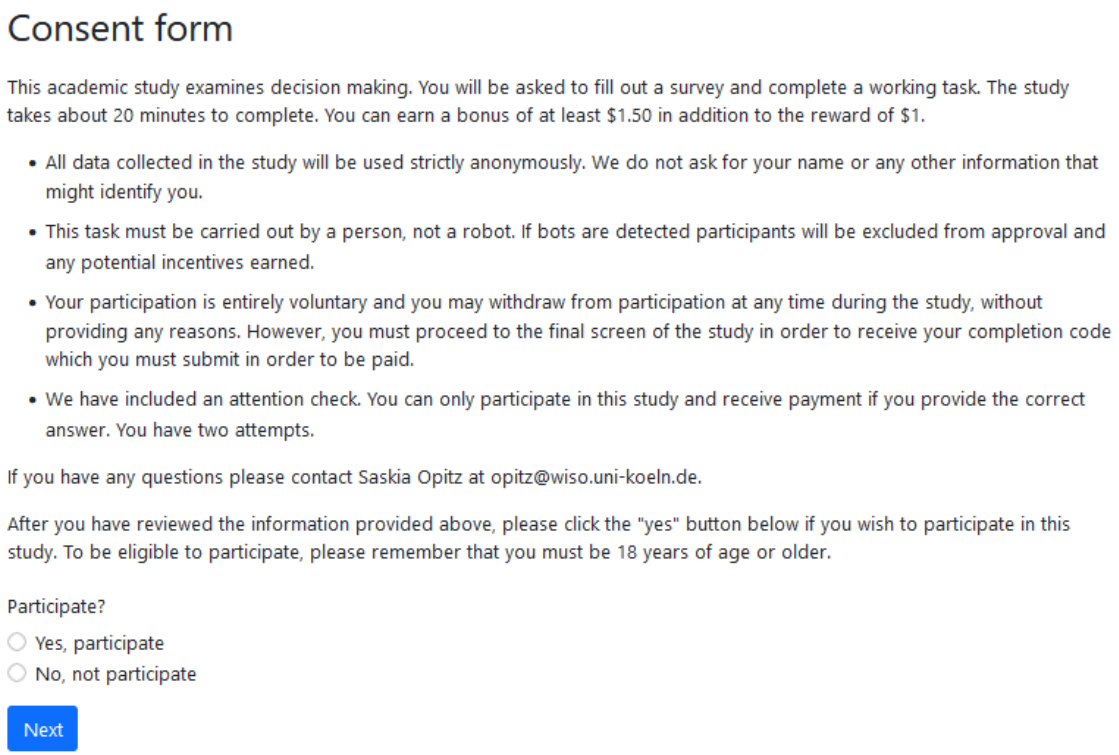


Figure A8: Consent to Participating in the Experiment

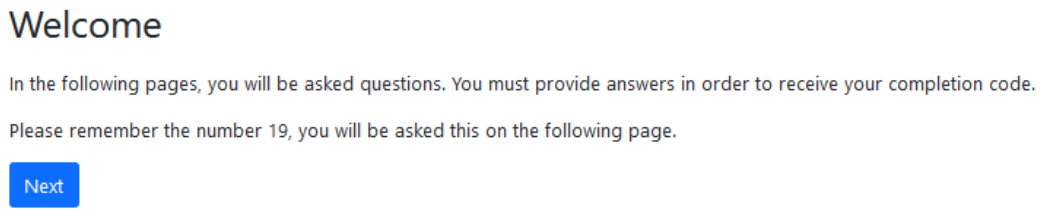


Figure A9: Welcome

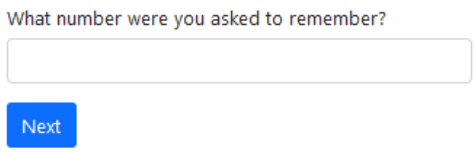


Figure A10: Attention Check

We will first ask you to provide some personal information and to evaluate several statements on the next pages. Please answer all of the following questions to the best of your ability.

The survey consists of multiple pages. **You will receive a bonus of $1.50 for completing the entire survey**. Additionally, you have the possibility to earn more money as a bonus in the course of this study.

Please note again that we do not ask for your name or any other information that might identify you.

Next

Figure A11: Survey Information

Survey

Please answer the following questions.

What is your age?

What is your gender?

- Male
- Female
- non-binary

Please indicate the highest level of education completed

- Less than High School
- High School or equivalent
- Vocational/Technical School (2 years)
- Some College
- College Graduate (4 years)
- Master's Degree (MA)
- Doctoral Degree (PhD)
- Other

Next

Figure A12: Demographics

Survey

Please answer the following questions. Please use a scale from 1 to 5, where 1 means you disagree strongly and 5 means you strongly agree. You can also use the values in-between to indicate where you fall on the scale.

I see myself as a person who...

	Disagree strongly 1	Disagree a little 2	Neither agree nor disagree 3	Agree a little 4	Strongly agree 5
...is reserved	○	○	○	○	○
...generally trusting	○	○	○	○	○
...tends to be lazy	○	○	○	○	○
...is relaxed, handles stress well	○	○	○	○	○
...has few artistic interests	○	○	○	○	○
...does things efficiently	○	○	○	○	○
...is outgoing, sociable	○	○	○	○	○
...tends to find fault with others	○	○	○	○	○
...does a thorough job	○	○	○	○	○
...gets nervous easily	○	○	○	○	○
...has an active imagination	○	○	○	○	○
...perseveres until the task is finished	○	○	○	○	○

Next

Figure A13: Big-5 Personality Traits

Survey

Please answer the following questions. Please use a scale from 1 to 5, where 1 means you disagree strongly and 5 means you strongly agree. You can also use the values in-between to indicate where you fall on the scale.

	Disagree strongly 1	Disagree a little 2	Neither agree nor disagree 3	Agree a little 4	Strongly agree 5
I always pay a lot of attention to how I do things compared with how others do things.	○	○	○	○	○
I am not the type of person who compares often with others.	○	○	○	○	○
I often compare how I am doing socially (e.g., social skills, popularity) with other people.	○	○	○	○	○
I see myself as someone who enjoys winning and hates losing.	○	○	○	○	○
I see myself as someone who enjoys competing, regardless of whether I win or lose.	○	○	○	○	○
I see myself as a competitive person.	○	○	○	○	○
Competition brings the best out of me.	○	○	○	○	○

Next

Figure A14: Social Comparison/Competitiveness

Survey

Please think about what you would do in the following situation. You are in an area you are not familiar with, and you realize that you lost your way. You ask a stranger for directions. The stranger offers to take you to your destination.

Helping you costs the stranger about **40 U.S. dollars** in total. However, the stranger says he or she does not want any money from you. You have six presents with you. The cheapest present costs **10 U.S. dollars**, the most expensive one costs **60 U.S. dollars**.

Do you give one of the presents to the stranger as a "thank you" gift?

- ○ No, would not give present
- ○ The present worth 10 U.S. dollars
- ○ The present worth 20 U.S. dollars
- ○ The present worth 30 U.S. dollars
- ○ The present worth 40 U.S. dollars
- ○ The present worth 50 U.S. dollars
- ○ The present worth 60 U.S. dollars

Next

Figure A15: Positive Reciprocity (Scenario)

Figure A16: Altruism (Scenario)

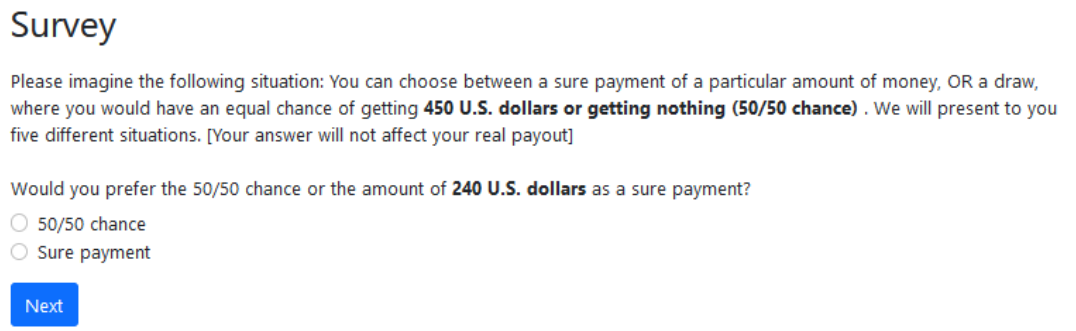


Figure A17: Risk Aversion (Staircase Measure)

Note: The following 4 screens encompassed the same instruction text and situations with different sure payment amounts.

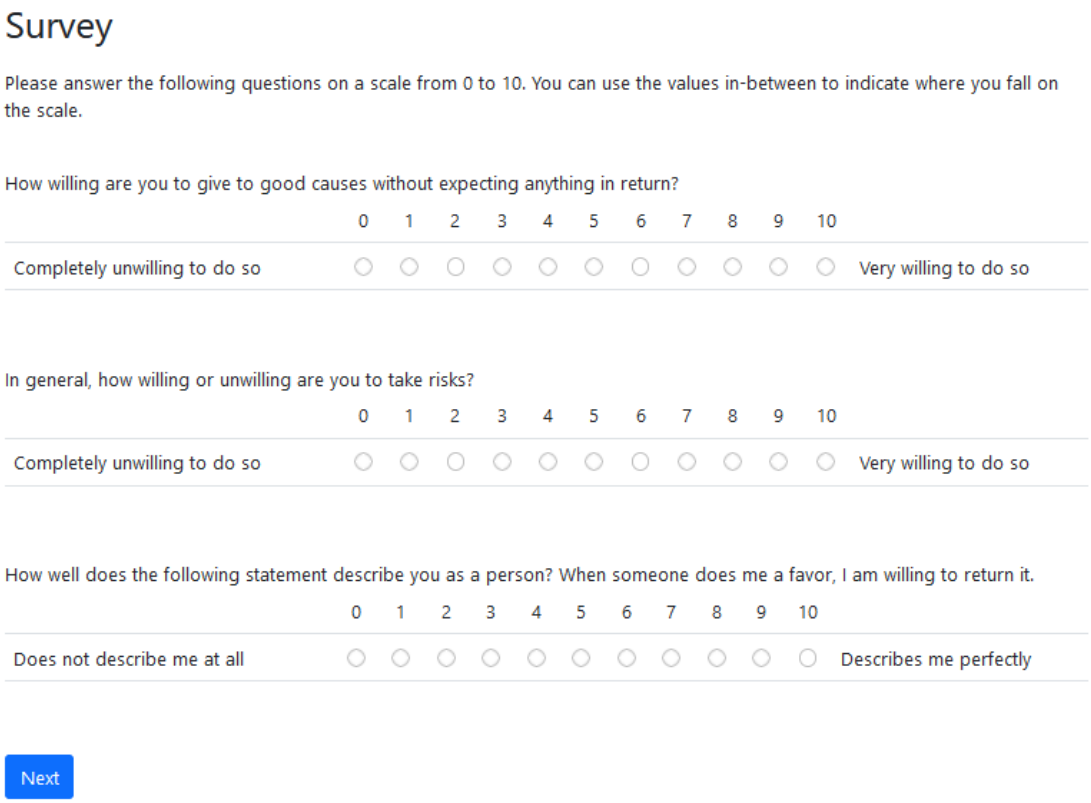


Figure A18: Positive Reciprocity/Risk Aversion/Altruism

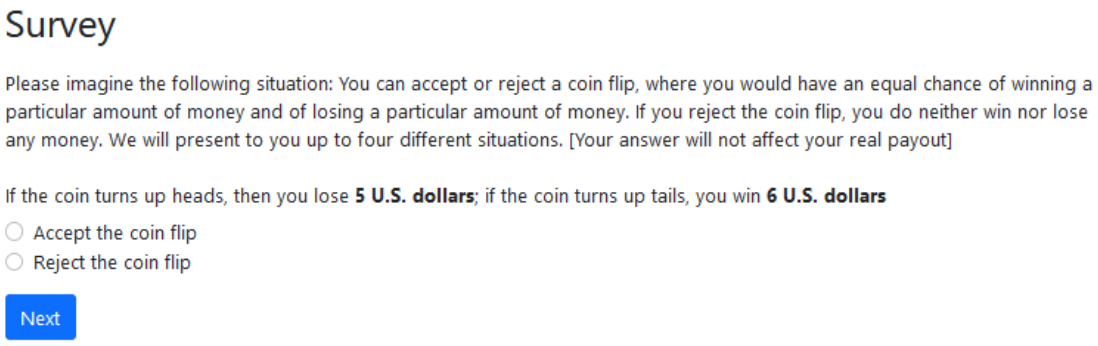


Figure A19: Loss Aversion

Note: Depending on the decisions, The following up to 3 screens encompassed the same instruction text and situations with different sure payment amounts.

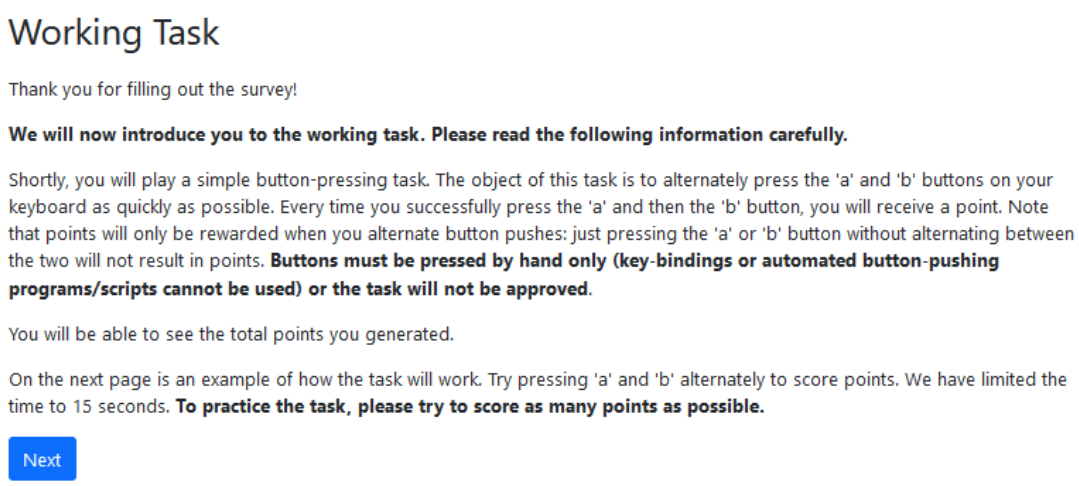


Figure A20: Introduction to the Working Task

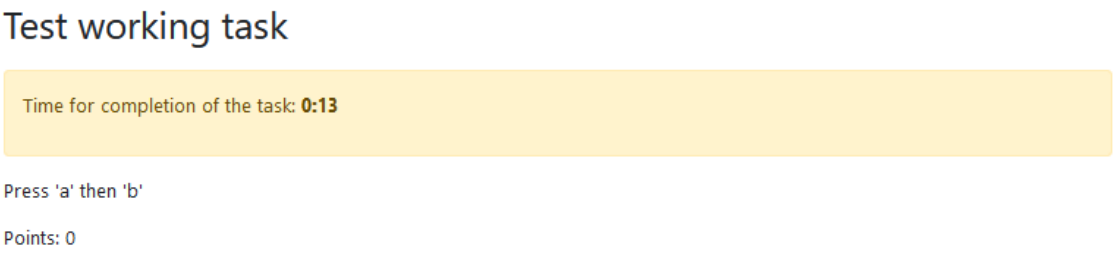


Figure A21: Test Phase (30 seconds)

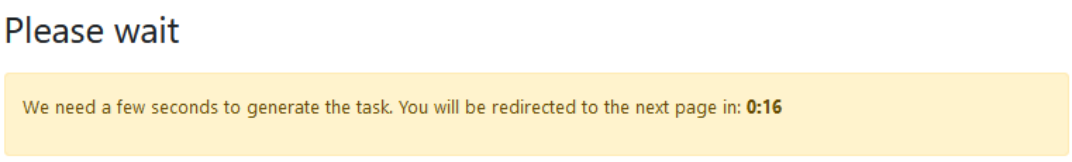


Figure A22: Waiting Screen (20 seconds)

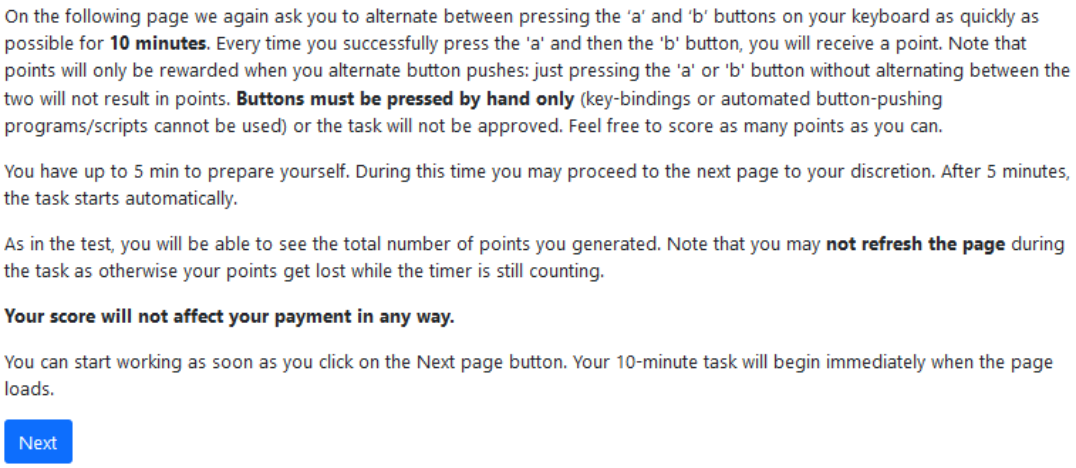


Figure A23: Treatment Information

Note: Treatment information for the different treatment groups:

Pay for Performance (PfP): "As a bonus, you will be paid an extra 5 cents for every 100 points that you score."

Bonus Gain: "As a bonus, you will be paid an extra $1 if you score at least 2000 points."

Gift & Goal: "Thank you for your participation in this study! In appreciation to you performing this task, you will be paid a bonus of $1. In return, we would appreciate if you try to score at least 2,000 points."

Bonus Loss: "As a bonus, you will be paid an extra $1. However, you will lose this bonus (it will not be placed in your account) unless you score at least 2,000 points."

Real-Time Rank Feedback: "You will receive a bonus that is based on how well you perform relative to others. On your work screen you will see how your current performance compares to that of others who previously performed the task. To that end you will see the percentage of participants who previously performed the task and whom you will outperform at your current speed. You will receive a bonus of $0.02 times the percentage of participants who performed worse than you at the end of the task. That is, you will for instance receive an additional bonus of $1.00 (=$0.02*50) if you perform better than 50% of the participants. The ranking shown on the screen is computed assuming you keep the speed with which you pressed 'a' and 'b' for the past 10 seconds. Your current percentile as well as your currently expected bonus is updated every 10 seconds." (As the text for this treatment is relatively long, the formatting differs slightly from the others. See Figure A24 for the formatting of this treatment information)

Social PfP: "As a bonus, you will be paid an extra 3 cents for every 100 points that you score. On top of that, 2 cents will go to Doctors Without Borders for every 100 points." (Figure A25 depicts a screen with more details on the donations workers could choose to look into)

Control: "Your score will not affect your payment in any way."

On the following page we again ask you to alternate between pressing the 'a' and 'b' buttons on your keyboard as quickly as possible for **10 minutes**. Every time you successfully press the 'a' and then the 'b' button, you will receive a point. Note that points will only be rewarded when you alternate button pushes: just pressing the 'a' or 'b' button without alternating between the two will not result in points. Buttons must be pressed by hand only (key-bindings or automated button-pushing programs/scripts cannot be used) or the task will not be approved. Feel free to score as many points as you can.

You have up to 5 min to prepare yourself. During this time you may proceed to the next page to your discretion. After 5 minutes, the task starts automatically.

As in the test, you will be able to see the total number of points you generated. Furthermore, you will see the bonus payments you generated. Note that you may **not refresh the page** during the task as otherwise your points get lost while the timer is still counting.

You will receive a bonus that is based on how well you perform relative to others. On your work screen you will see how your current performance compares to that of others who previously performed the task. To that end you will see the percentage of participants who previously performed the task and whom you will outperform at your current speed.

You will receive a bonus of $0.02 times the percentage of participants who performed worse than you at the end of the task. That is, you will for instance receive an additional bonus of $1.00 (=$0.02*50) if you perform better than 50% of the participants. The ranking shown on the screen is computed assuming you keep the speed with which you pressed 'a' and 'b' for the past 10 seconds. Your current percentile as well as your currently expected bonus is updated every 10 seconds.

You can start working as soon as you click on the Next page button. Your 10-minute task will begin immediately when the page loads.

Next

Figure A24: Treatment Information (RTR Feedback Treatment)

Donation Details

Who is receiving the donations?

Doctors Without Borders (link)

What is the timeline for these donations?

After the study is complete, we will take a few days to verify that all participants obeyed the rules of the study and played fairly. Then we will calculate the total amount of donations earned and donate it directly to Doctors Without Borders. We will upload the donation receipt here.

Additional questions

If you have additional questions regarding this process, please contact me at opitz@wiso.uni-koeln.de.

Kind regards,

Saskia Opitz

On behalf of Dirk Sliwka

Go back to Instructions

Please remember that after 5 minutes, the task starts automatically.

Figure A25: Donation Details (Social PFP Treatment)

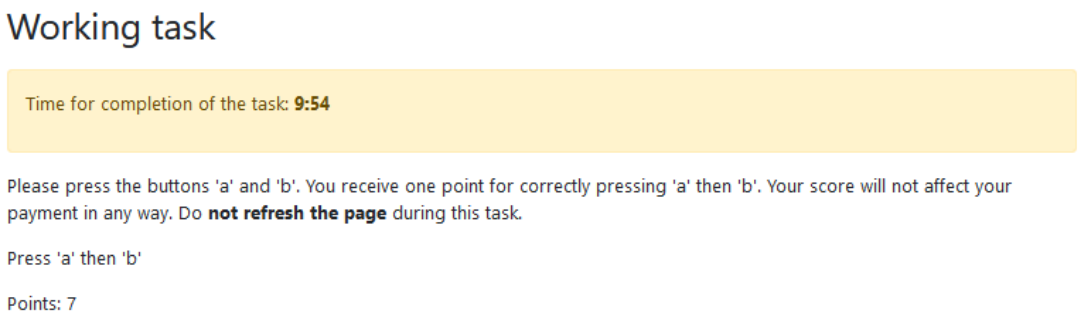


Figure A26: Working Stage (10 minutes)

Note: Depending on the treatment, this screen also entailed information on the current bonus amount, current bonus amount and rank, or current bonus and donation amount.

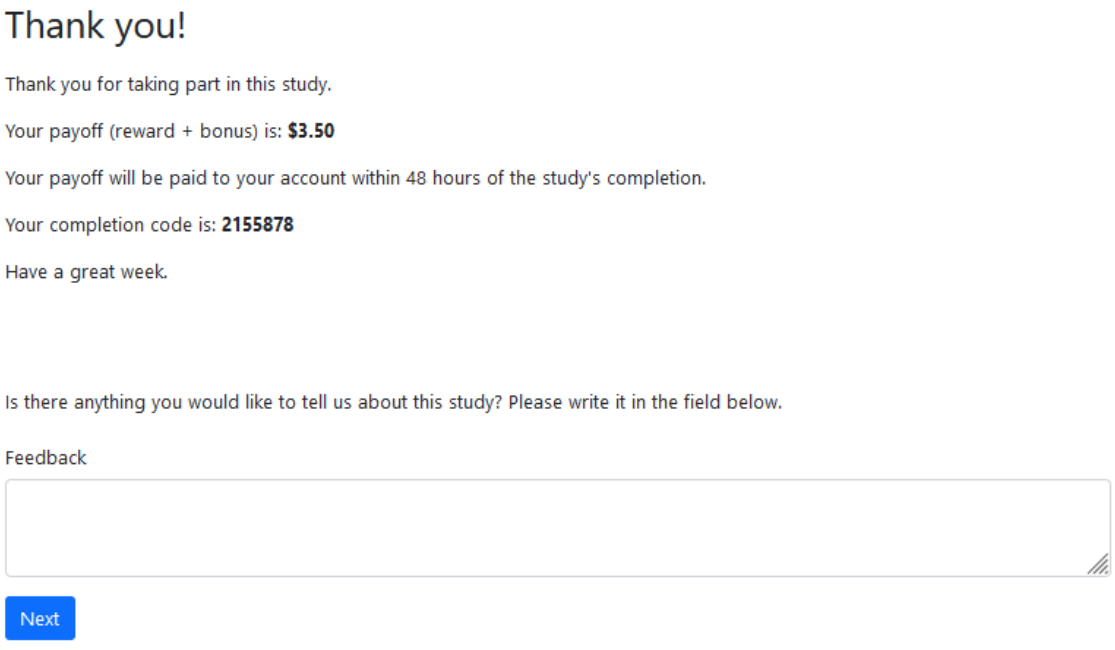


Figure A27: End Screen

Note: Depending on the treatment, this screen also entailed information on the final rank or final donation amount.