ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 187

The Targeted Assignment of Incentive Schemes

Saskia Opitz
Timo Vogelsang
Dirk Sliwka
Tom Zimmermann

August 2022

www.econtribute.de

UNIVERSITÄT BONN
Universität zu Köln

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/1-390838866 is gratefully acknowledged.

The Targeted Assignment of Incentive Schemes*

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

This Version: April 8, 2022

Abstract

A central question in designing optimal policies concerns the assignment of individuals with different observable characteristics to different treatments. We study this question in the context of increasing workers' performance by using targeted incentives based on measurable worker characteristics. To do so, we ran two large-scale experiments. The key results are that (i) performance can be predicted by accurately measured personality traits, (ii) a machine learning algorithm can detect such heterogeneity in worker responses to different schemes, and (iii) a targeted assignment of schemes to individual workers increases performance in a second experiment significantly above the level achieved by the single best scheme.

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

JEL classification: C21, C93, M52

*We thank Stefano DellaVigna, Jonathan de Quidt and Matthias Heinz as well as participants of the Young ECONtribute Program seminar 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 Germany´s 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 workers, employers can choose from a variety of different incentive schemes. Previous research has already shown positive average performance effects, for instance, for performance pay (e.g., Lazear 2000; Bandiera, Barankay and Rasul 2007), 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).[1] But while one person may perform best under e.g. a performance pay scheme, others may be motivated more effectively through different types of incentives. This paper investigates whether and to what extent worker performance can be improved by the targeted assignment of incentive schemes based on individual worker characteristics.

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

To study the potential of targeted incentive schemes for performance improvements, we ran two consecutive large-scale real-effort experiments with altogether around 12,000 workers on Amazon MTurk. In both experiments, we hired workers for a real-effort task developed by DellaVigna and Pope (2018). The key question is, therefore, whether (i) a machine learning algorithm trained with data on individual characteristics can detect heterogeneous responses, 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 a first step, we ran an initial experiment implementing 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). The schemes 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,

[1] Please note that due to the large number of existing studies this list does not claim to be complete. Moreover, this list is not an evaluation of the significance of individual studies. For reasons of clarity, we limit ourselves to isolated examples. For a more complete overview see Bandiera, Barankay and Rasul (2011), Sprinkle and Williamson (2006), Lazear (2018).

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[2].

The highest average performance in the first experiment is achieved by the *Bonus Loss* scheme that awards a bonus that is lost when the worker does not achieve a specific target value.[3] But estimating conditional average treatment effects, we detect significant heterogeneity based on worker characteristics in the data from experiment 1. In other words, our estimated model predicts that for subsets of workers of different characteristics, different schemes would lead to a higher performance.

We then validated this prediction in a second experiment, where we again elicited the respective workers' characteristics. In this second experiment, we compare three treatments (i) A control treatment where all workers work under a fixed wage, (ii) a *Best ATE* treatment where all workers work under the scheme that generated the highest average treatment effect in the first experiment, and (iii) an *Algorithm* treatment where workers are exposed to the scheme that is predicted to yield the highest performance conditional on the specific characteristics of each individual worker.

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 set resulting from this procedure comprises 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.

[2]In particular, we measure the Big-5 personality traits (Benet-Martínez and John 1998; John, Donahue and Kentle 1991; John, Naumann and Soto 2008; Rammstedt and John 2007), risk preferences (Falk et al. 2016, 2018), altruism (Falk et al. 2016, 2018), positive reciprocity (Falk et al. 2016, 2018), loss aversion (Gächter, Johnson and Herrmann 2021), competitiveness (Fallucchi, Nosenzo and Reuben 2020), social comparison (Gibbons and Buunk 1999)

[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 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, Hoffman and Moser 2005; Armantier and Boly 2015; Imas, Sadoff and Samek 2017; Van der Stede, Wu and Wu 2020). Others do not find a statistically significant difference (e.g., Grolleau, Kocher and Sutan 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 (2021) for a meta-analysis.

We find that the treatment with the targeted scheme assignment significantly outperforms the loss treatment that achieved the highest treatment effect in Experiment 1: While the loss treatment raises performance by 23.9% over the level of the fixed wage control group, the targeted assignment raises performance by 29.3% and thus leads to a 5.4 percentage points or 22.5% higher treatment effect.

The algorithm assigns the incentive scheme based on information from a survey on the subjects' personality traits as well as social and economic preferences. The next question we address is, thus, whether and to what extent the quality of the assignment procedure depends on the reliability of the elicited traits. As traits are self-reported subjects may, for instance, differ in their diligence filling out the survey. For one, more diligent survey responses should lead to more precise assessment of traits and thus to better assignment. Yet the algorithm could also pick up patterns in the survey responses that predict this diligence and at the same time are informative for optimal incentive assignment.[4] To study this, we create a measure of a subjects' reliability by studying the consistency in the answers to items measuring the same trait. We find that reliability of survey responses strongly affects the quality of the assignment procedures as the effect of the *Algorithm* treatment is the larger, the more reliable the survey responses are. Hence, the contribution of mere pattern recognition in responses does not appear to be large and rather precisely elicited traits are crucial for the performance of the assignment algorithm.

As experiment 2 comprises both newly hired workers and workers who also took part in experiment 1, we can also compare the performance effects of the *Algorithm* treatment in these two sub-samples. We find that the performance gains are substantially larger in the sub-sample of workers who already had taken part in the first experiment. This is the case even though the algorithm did not use information on their identity in the first experiment in the assignment procedure. Comparing these sub-samples in more detail, we find that the sample of "new hires" is significantly different from the "retakers" in some observable characteristics. Moreover, survey responses in the sample of new hires exhibit a significantly lower reliability on average. We then show that there are sizeable treatment effects in the subset of new hires with reliable answering patterns. Hence, an important precondition for a sensible targeted assignment of incentive schemes is a reliable measurement of the subjects' traits.

[4]For instance, when a subject always ticks the first box in the survey this will lead to completely unreliable direct measures for each trait, but the pattern in itself may be indirectly revealing about the subject's traits and this could be picked up by the algorithm.

Our study adds to different strands of the literature. We contribute to the literature on heterogeneous effects of incentives schemes. Several studies have found effect heterogeneity with respect to factors such as, for example, gender (Gneezy, Niederle and Rustichini, 2003; Niederle and Vesterlund, 2007; Delfgaauw et al., 2013), social preferences (Bandiera, Barankay and Rasul 2005), task motivation (Ashraf, Bandiera and Jack, 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, Sliwka and Vogelsang 2021). 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.

Our findings also relate to the literature on sorting into incentive schemes. Several studies have shown, that individuals sort by their preferences when choosing between incentive schemes (Lazear, 2000; Banker et al., 2000; Cadsby, Song and Tapon, 2007; Dohmen and Falk, 2011; Larkin and Leider, 2012). But 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, we complement a growing literature that uses machine learning methods to estimate heterogeneous treatment effects and then uses such estimates for optimal policy assignment.[5] Several studies provide parametric (Imai and Ratkovic, 2013) and non-parametric (Athey and Imbens, 2016; Wager and Athey, 2018; Farrell, Liang and Misra, 2021) estimators to identify subgroups with high expected treatment effects while taking the issue of multiple hypothesis testing into account. We compare several of those estimators and find 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, one can proceed to define optimal policy assignments (Hirano and Porter, 2009; Hitsch and Misra, 2018; Kitagawa and Tetenov, 2018; Caria et al., 2020). Such estimated policy assignments are typically used to describe observational data (Kleinberg et al., 2015, 2017), but are not validated out-of-sample (with Dubé and Misra (forthcoming) being an exception in the context of personalized pricing).

The paper proceeds as follows. First we present the design and results of Experiment 1 in section 2. Then we explain the implemented algorithm and the resulting assignment procedure in section 3 and report the results of experiment 2 in section 4. Section 5 provides further analyses and robustness checks and section 6 concludes.

[5] See Athey and Imbens (2017) and Athey and Imbens (2019) for comprehensive overviews.

2 Experiment 1

2.1 Experimental Design

The first experiment consists of two parts. First, workers have to complete a survey eliciting 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)[6]. In the second part, workers work on a real-effort task. We use the real-effort task developed by DellaVigna and Pope (2018). In this task, workers have to repeatedly press the 'a' and the 'b' button on their keyboards to score points. One point is awarded for each time they correctly press first 'a' then 'b'[7]. 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 ability in such type of tasks[8]. After the test phase and a short waiting screen[9], workers receive the 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.[10] 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 gain (*Bonus Gain*) or loss (*Bonus Loss*), respectively. Additionally, we include three treatments which are similar to the ones by DellaVigna and Pope (2018) but are adjusted to make them more comparable to the other three treatments in (i) the payments made, (ii) the bonus workers can reach for themselves, and (iii) guidance provided on how many points should be reached. 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.

[6] See Appendix A for the list of characteristics and scale references. Note that participants cannot skip questions, but they can withdraw from the study at any time.

[7] For a screenshot of the working stage for the *Social PfP* treatment see Figure 4 in Appendix A

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

[9] We included the waiting screen in experiment 1 so that the sequence of screens did not differ between experiment 1 and experiment 2. In experiment 2, the waiting time was necessary to allocate the participants in the *Algorithm* treatment to their predictably best treatment (see Section 4.1)

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

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 and the completion code, which they need to submit the task for payment.

2.2 Experimental Procedure

We implemented the experiment using oTree (Chen, Schonger and Wickens 2016). Workers are invited via MTurk.[11] 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[12]). Furthermore, they are asked for their consent to participate in the study from which they know they can withdraw at any time.

The experiment ran for 2.5 weeks in September 2021. We required workers to be located in the US.[13] In total, more than 6,649 workers submitted the task for payment. Based on pre-registered criteria[14] we excluded 584 workers resulting in a final sample consisting of 6,065 workers.[15]

[11] Evidence suggests that MTurk findings are generally similar to findings in laboratory or field settings (Horton, Rand and Zeckhauser 2011; Farrell, Grenier and Leiby 2017; Snowberg and Yariv 2021).

[12] 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 also the workers in the control group 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.

[13] 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

[14] 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 will collect data as well.

[15] 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* mainly comes 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.

Table 1: Treatments

Treatment	Incentive Scheme Text
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.
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.
Control	Your score will not affect your payment in any way.

The average duration of the experiment was around 19 minutes (median duration around 17 minutes) and mean payoff was $3.32 ($10.60 per hour; $11.93 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 9 in Appendix B.

The following stratified randomization procedure is applied to achieve balanced sampling into the treatments: Strata are 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 these strata, treatments 1 to 7 are assigned in a random order such that in each stratum each treatment is 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.[16]

[16]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.

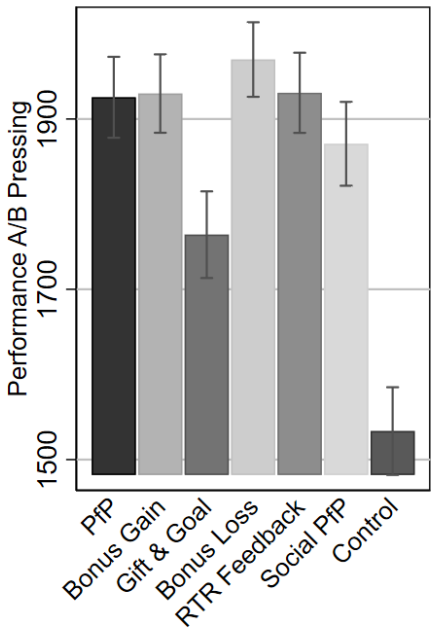


Figure 1: Results of Experiment 1

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. Vertical lines correspond to the 95% confidence interval. For corresponding regression results, see Table 11 in Appendix B.

3 Heterogeneity in Treatment Effects

To gain insight into the effect of incentive schemes beyond the average treatment effects analyzed in section 2, we estimate conditional average treatment effects (CATEs) defined in equation (1).

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

In equation (1), the CATE is 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 CATE is the same for all individuals and the same as the average treatment effect.

We compare 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, Liang and Misra 2021), Indirect Random Forests (Breiman 2001; Foster, Taylor and Ruberg 2011) as well as a Doubly Robust approach (Chernozhukov et al. 2018*a*).[17]

To select the best-performing algorithm in our context, we analyzed the results of experiment 1 using the method in Hitsch and Misra (2018). In particular, we train each algorithm on parts of the sample and predict CATEs out-of-sample using cross-validation. We compare the performance of each algorithm 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.

The algorithm yielding the highest performance was the indirect random forest approach.[18] Based on this finding, we proceed to use the indirect random forest approach to estimate CATEs in the remainder of the study.[19]

Using the indirect random forest approach proceeds in two steps. We follow the two steps for each of our treatments separately. In step 1, we train two random forests, one to predict the effort for the treatment group using the personal characteristics elicited by the survey as features, one to predict the effort for the control group using the same features. Using the estimated models, we predict 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 use 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 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 compare the performance of the algorithm when restricting the number of potential

[17]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.

[18]Table 10 in Appendix B shows the performance for each of the algorithms.

[19]While indirect random forests turns 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, Liang and Misra 2021).

incentive schemes 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.[20]

To assess the quality of the algorithmic assignment, we conduct the following exercise: For each group of workers with the same predicted assignment, we compare their performance across the incentive schemes that they were actually assigned to in experiment 1. Table 2 shows results. In column (1), we restrict the sample to workers that the algorithm would have assigned to the *Bonus Loss* scheme. Looking at the performance of those workers across actually assigned schemes, we observe the highest performance gain for the workers that 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 restrict the sample to workers that the algorithm would have assigned to *RTR Feedback* or *Social PfP*, respectively, we observe 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.[21]

The importance of specific traits for assigning different schemes can be illustrated in partial dependence plots. For each estimated treatment algorithm, we, for instance, can depict how a change in a particular covariate affects the predicted performance.[22] 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.

[20] Figures 5-7 in Appendix B plot the feature importance for the remaining features. Across all trained models, age and altruism are typically among the most predictive predictor variables. Otherwise, most predictive variables vary depending on the incentive scheme.

[21] Panel (a) in Figure 8 in Appendix B 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.

[22] 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, Tibshirani and Friedman (2009) for details.

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=RTR Feedback	0.000	0.005	0.530
P-value Bonus Loss=Social PfP	0.000	0.907	0.081
P-value RTR Feedback=Social PfP	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.

Figure 2 shows an example.[23] 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.[24]

[23] See Figures 9 and 10 in Appendix B for a full set of partial dependence comparisons.

[24] 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 assignment from bonus-loss to *Social PfP*, which cannot be reflected in the simple ceteris paribus comparison of Figure 2.

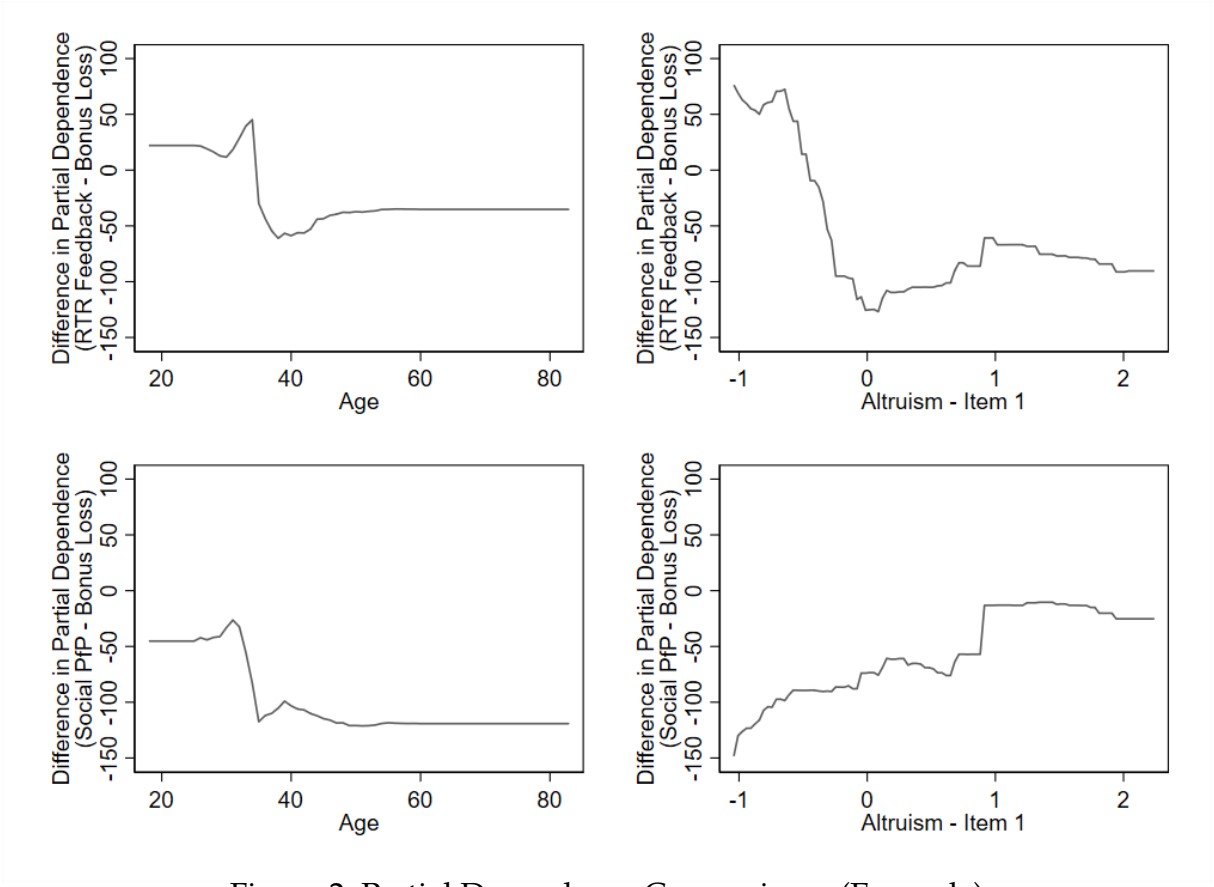


Figure 2: Partial Dependence Comparisons (Example)

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 9 and Figure 10 in Appendix B for further partial dependence plots.

4 Experiment 2

4.1 Experimental Design

In the second experiment, we first again elicit workers' characteristics and provide instructions following the same protocol as experiment 1.[25]

After the survey, workers again receive instructions on the 'a/b'-pressing task and can test the task for 30 seconds. We again use the points scored in this test phase as a proxy for ability in this task. After the test, all participants see a waiting screen for 20 seconds. The waiting screen was necessary as during this time, the participants in the *Algorithm*

[25] A list of the elicited characteristics can be seen in Appendix A.

treatment are assigned to their incentive schemes. On the following screen, all workers again receive the instructions for the real-effort task as well as additional information on their respective incentive scheme (see Table 1 for the exact wording of the incentive schemes). 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, i.e., the *Loss* scheme. 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 was similar to that of experiment 1. The experiment ran for 3 weeks in November 2021. We again required workers to be located in the US.[26] In the first two weeks, we recruited only workers who had not taken part in experiment 1. After that, we dropped this restriction. Due to the sequential randomization procedure, treatment shares are balanced in both populations.[27] 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.[28] 4,282 were "new hires", and 2,096 retook the study after already having completed experiment 1. The sample size of the second experiment is based on a power analysis conducted after the first. To be specific, we used the method in Hitsch and Misra (2018) to predict how large the treatment effect of the *Algorithm* treatment will be in experiment 2. Based on this predicted effect size, we performed a power analysis in particular for the comparison of the *Algorithm* and *Best ATE* treatments. Based on this, we implemented a sample size of 6,200 workers (3,000 for each treatment group and 200 for the control group).

[26] 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

[27] We systematically compare the treatment effects within these groups in section 5.4

[28] 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.

The average duration of the experiment was around 19 minutes (median duration around 17 minutes) and mean payoff was \$3.30 (\$10.54 per hour; \$11.80 per hour median).[29] 49.3% of workers in experiment 2 indicated that they were female. 73% had at least a college degree and the mean age was around 39 years. Again, our MTurk sample somewhat over-represents younger and more educated groups of the U.S. population. Descriptive statistics are shown in table 9 in the Appendix B.

The assignment of the participants to the treatments is determined as follows. First, workers are randomly assigned to the first control group (i.e., no incentive scheme) or to receiving an incentive scheme.[30] For the workers who receive an incentive scheme, we constructed strata based on the entry time of the workers to the study, i.e., the first two workers to click on the link and thus enter the study belong to one stratum, the two workers entering afterwards belong 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 one individual to the treatment suggested by the algorithm.

4.3 Incentive Scheme Assignment

Figure 3 shows the share of workers in the Algorithm treatment of experiment 2 that are assigned to the three remaining schemes based on their elicited characteristics. While the algorithm still assigns about 39.25% of the subjects to the *Bonus Loss* scheme, a higher share of about 48.01% is assigned to the *Real-time Rank Feedback* condition and a smaller share of 12.75% to the *Social PfP* scheme.

4.4 Results Experiment 2

We now address the question whether, and if yes to what extend the algorithmic assignment of the scheme can improve performance above the level of the scheme that achieved the highest average treatment effect.

[29] 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.

[30] 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 3000 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.

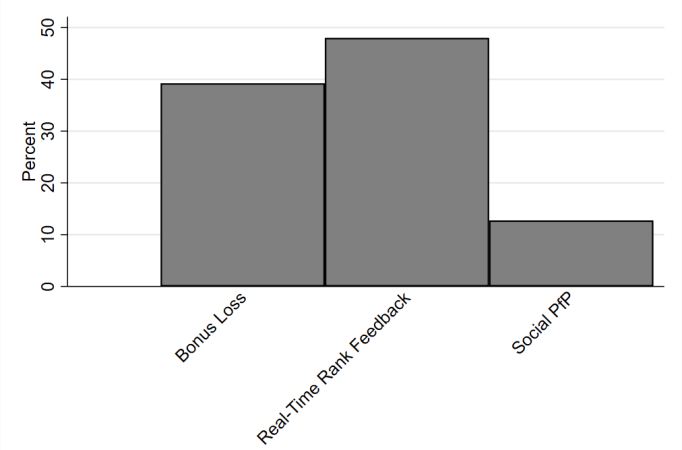


Figure 3: Share of Workers by Incentive Scheme

Note: This figure shows the share of workers assigned to each of three incentive schemes in the algorithm treatment. The algorithm treatment is based on predicted conditional average treatment effects, and each individual is assigned to the incentive scheme with the highest predicted individual treatment effect. See section 3 for details.

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).[31] The $Algorithm_i$ dummy indicates an observation from the treatment where the assignment is based on the algorithm. In Columns (3) and (4), we restrict the sample to workers in one of the two treatment groups so that we can directly compare their performance.

Columns (1) and (3) report results of an OLS regression with batch fixed effects as well as the ability proxy as control variables, and columns (2) and (4) includes demographics i.e., age, gender, and education level, as further control variables.

As Table 3 shows, the $Best\ ATE_i$ treatment raises performance above the level of the fixed wage control group by about 23.9%. The effect of the $Algorithm_i$ treatment is 29.3%.[32] The targeted assignment of incentive schemes thus significantly increases the overall incentive effect by 5.4 percentage points or 22.5% ($p = 0.013$ and $p = 0.019$ respectively for columns (1) and (2)). This corresponds to a more than 4% increase in performance in comparison to the group working under the single best incentive scheme in experiment 1.

[31] See Table 12 in Appendix B for the results using absolute performance.

[32] This is based on the results from column (1) in Table 3. Note that $exp(0.214) = 1.239$ and $exp(0.257) = 1.293$.

Table 3: Main Results: Effect on Performance

	$log(Performance)_i$			
	(1)	(2)	(3)	(4)
$Algorithm_i$	0.257***	0.257***	0.043**	0.041**
	(0.057)	(0.058)	(0.017)	(0.017)
$Best\ ATE_i$	0.214***	0.216***		
	(0.058)	(0.058)		
P-value Best ATE=Algorithm	0.013	0.019		
Reference Group	Control	Control	Best ATE	Best ATE
Dem. Controls	No	Yes	No	Yes
Observations	6,377	6,377	6,147	6,147
Adjusted R-squared	0.112	0.118	0.110	0.117

Note: In this table, we report the results of regressions of log(Performance) on treatment dummies for the *Best ATE* treatment as well as the *Algorithm* treatment. In columns (3) and (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 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. In columns (2) and (4), we further include demographic controls, i.e. age, gender dummies as well as education level dummies, as controls. Standard errors are clustered at the batch level, and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

5 Further Analyses

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.[33] Within each of these sub-samples we estimate the average treatment effect of the respective scheme assigned by the algorithm in comparison to the $Best\ ATE_i$ treatment (i.e. the *Bonus Loss* scheme). Results are displayed in Table 4. The first sub-sample comprises all subjects from the three treatments for which the algorithm predicted that their performance is highest under the *Loss* scheme. Note that here the $Best\ ATE_i$ and the $Algorithm_i$ treatments implement exactly the same scheme on a subsample selected by exactly the same procedure and thus both point estimates have the same magnitude.

[33] That is the observations from the $Best\ ATE_i$ treatment are allocated to the subsample associated to the scheme that the algorithm would have assigned them to.

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.[34]

Table 4: Effects in Sub-Samples

	$log(Performance)_i$		
	Predicted Bonus Loss	Predicted RTR Feedback	Predicted Social PfP
	(1)	(2)	(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 characteristic. 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.

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

5.2 Assignment Group Characteristics

As the algorithm makes use of the rich information contained in the different patterns of survey response behavior and potentially complex interaction structures, it is impossible to depict the specific functional form used. Nevertheless, it is instructive to consider which of the measured traits are directly associated with the likelihood that a person is assigned to a specific scheme.

[34] Note that $exp(0.069) = 1.071$.

To illustrate this, we estimated 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.

Note that all preference and personality trait measures are standardized such that we can compare the magnitudes of the respective regression coefficients. Several features stand out. Older workers and women 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.[35] Also in line with previous research that has shown that women perform less well under competitive incentives (see e.g. Gneezy, Niederle and Rustichini 2003), women are less likely to be assigned to the *Real-time Rank Feedback* scheme. Somewhat surprisingly, women are also less likely assigned to the *Social PfP*.[36]

Among the trait and personality measures, we observe the most pronounced differences with respect to altruism. In line with straightforward reasoning, more altruistic subjects 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 which are all associated with prosocial traits also are positively associated with the likelihood to be 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. 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.

5.3 Reliably Measured Traits or Pattern Recognition?

The estimated assignment algorithm relies on survey responses to assign individuals to incentive schemes. A natural question that arises is whether participants' survey responses are informative in the sense that they provide information about the personality traits that they are supposed to elicit. A different possibility would be that the mere pattern of responses, perhaps unconsciously, provides information that is useful for the

[35] 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.

[36] 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.

Table 5: Group Characteristics (Logit)

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

assignment. For example, we observe a number of participants who always click on the right-most column throughout the survey which is suggestive of mere pattern clicking rather than providing informative responses on traits (since we reverse-coded some of the survey items).

To investigate this question in more detail, we generate a consistency measure of the responses to different survey items measuring the same trait. To this end, we can make use of the fact that several of the psychological scales we use include reverse-coded items.[37] As consistency measure of survey answers, we use 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).

If it is indeed important that subjects respond to the survey questions in a consistent manner and the assessed traits are indeed crucial for assignment, we should observe that the *Algorithm* treatment performs substantially better for subjects who score high in consistency. If, however, mere pattern recognition drives the treatment effects, even inconsistent answers may help as the fact that the self-assessments appear inconsistent are informative per se.

To investigate this question, we study whether and to what extent the size of the treatment difference between the *Algorithm* and the *Best ATE* treatment depends on the consistency of the survey answers. In particular, we regress log performance on the measure for the consistency of individual survey answers, on a dummy for being in the *Algorithm* treatment, and the interaction between both on a sample comprising the data from the *Algorithm* and the *Best ATE* treatments. The results are reported in Table 6.

As the regression results show, the treatment effect is larger, the larger the consistency of the survey responses. A by one standard deviation higher consistency is associated with a treatment effect that is twice as large as the effect at the sample average. By the same token, the treatment effect vanishes for survey respondents with a one standard deviation lower consistency.

These results indicate that a reliable measurement of traits improves the quality of the assignment substantially. Or, in other words, it seems unlikely that mere pattern recognition in the survey responses drives the value of the targeted assignment through the algorithm.

[37] This applies to the following traits: conscientiousness, agreeableness, emotional stability, extraversion, openness, and social comparison. An example is for instance the conscientiousness scale which includes for instance 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.*"

Table 6: Effect of the Algorithm Depending on Consistency of Responses

	$log(Performance)_i$	
	(1)	(2)
$Algorithm_i$	0.042**	0.041**
	(0.017)	(0.017)
$Consistency_i \times Algorithm_i$	0.049**	0.049**
	(0.020)	(0.019)
$Consistency_i$	0.039	0.022
	(0.024)	(0.024)
Reference Group	Best ATE	Best ATE
Dem. Controls	No	Yes
Observations	6,147	6,147
Adjusted R-squared	0.115	0.120

Note: In this table, we report the results of regressions of log(Performance) on a measure for the consistency of individual survey answers, on a dummy for being in the *Algorithm* treatment, and the interaction between both. 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 exclude the control group so that the Best ATE treatment group is the reference group for the Algorithm dummy. We include batch fixed effects as well as an ability proxy as controls. In column (2), we further add demographics, i.e. age, gender dummies and education dummies, 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.

5.4 The Role of Sample Characteristics

Experiment 2 comprises both, newly hired workers and workers who had also taken part in experiment 1 before. Comparing the performance of the algorithm treatment in these two groups of workers provides an opportunity to test how the algorithm performs on a sample of workers on which the algorithm had not been trained (i.e. who were not part of experiment 1).

To analyze this question we pool the data from the *Algorithm* and *Best ATE* treatments in experiment 2 and regress log performance on an *Algorithm* treatment dummy interacted with a dummy indicating whether a person is a *New Hire*. Table 7 shows the respective regression results. We find that the achieved performance gains are substantially larger in the sub-sample of workers who already had taken part in the first experiment. Here *Algorithm* outperforms *Best ATE* by more than 9.4%. This is the case even though the algorithm did not use the information on their identity in the first experiment in the assignment procedure for the scheme in experiment 2. However, we also find that the algorithm hardly outperforms the *Best ATE* treatment in the group of newly hired workers as the respective interaction term shows.

Table 7: Comparison of New Hires and Retakers

	$log(Performance)_i$	
	(1)	(2)
$Algorithm_i$	0.096***	0.094***
	(0.028)	(0.029)
$New\ Hire_i \times Algorithm_i$	-0.079*	-0.080*
	(0.040)	(0.041)
$New\ Hire_i$	0.127***	0.147***
	(0.046)	(0.045)
Reference Group	Best ATE	Best ATE
Dem. Controls	No	Yes
Observations	6,147	6,147
Adjusted R-squared	0.111	0.118

Note: In this table, we report the results of regressions of log(Performance) on a dummy for being a *New Hire*, on a dummy for being in the *Algorithm* treatment, and the interaction between both. We exclude the control group so that the Best ATE treatment group is the reference group for the Algorithm dummy. We include batch fixed effects as well as an ability proxy as controls. In column (2), we further add demographics, i.e. age, gender dummies and education dummies, 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.

To understand this difference in the algorithm's performance, we compare the observable characteristics in both groups. We find that the two samples are significantly different from each other with respect to several characteristics. A particularly striking difference is that new hires are less consistent in their answering behavior in the survey: As Table 14 in Appendix B shows, consistency is significantly larger for those workers who participate a second time in the experiment than for newly hired workers (column 1).[38] Moreover, as column 2 shows, these workers had already exhibited a significantly higher consistency in experiment 1 than workers who only took part in the first experiment but not in the second. Hence, this effect is unlikely to reflect learning but rather points towards differences in selection. That is, the sample of workers who took part twice is different with respect to the workers' inherent characteristics.

[38] Table 13 reported in Appendix B shows that there are further significant differences in average age, ability, level of education.

Importantly, as we have shown in section 5.3, it is crucial for the treatment effect of the targeted assignment that the traits measured in the survey are consistently measured. We thus use the consistency measure from section 5.3 and restrict the sample to subsets of new hires that score high on the consistency measure. That is, we rank the workers according to the consistency measure and consider only new hires with a consistency in survey responses larger than a specific percentile.

Table 8: Effect Depending on Answer Consistency: New Hires

	$log(Performance)_i$		
	Consistency ≥ 40% (1)	Consistency ≥ 50% (2)	Consistency ≥ 60% (3)
$Algorithm_i$	0.023	0.049*	0.061**
	(0.023)	(0.026)	(0.029)
Reference Group	Best ATE	Best ATE	Best ATE
Dem. Controls	Yes	Yes	Yes
Observations	2,423	2,007	1,548
Adjusted R-squared	0.104	0.100	0.107

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 restrict the sample to new hires with a consistency in survey responses larger than or equal to the 40th percentile of new hires. In column (2) and column (3), we restrict the samples to new hires with a consistency in survey answers larger than or equal to the 50th percentile and 60th percentile, 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, an ability proxy as well as demographics, i.e. age, gender dummies and education dummies, 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.

Comparing the effectiveness of the algorithm only among new hires for different consistency thresholds, Table 8 shows that the targeted assignment can perform well also for newly hired workers – but only if their answers in the survey are consistently measured. It does not lead to significant performance increases if this is not the case. This again underscores the importance of a reliable measurement of traits, in particular, when the assignment procedure is applied to samples with different inherent characteristics.

6 Conclusion

We show that workers' productivities are strongly affected by different incentive schemes. More importantly, we find that the productivity response can be predicted by accurately measured personality traits, and that recently advanced methods that combine machine learning with causal inference can detect such heterogeneous responses to incentive schemes. Moreover, the targeted assignment of incentive schemes based on individual worker characteristics can increase overall worker performance above the level achieved by a scheme that performs best on average.

In particular, the targeted assignment of three schemes, a loss-framed bonus scheme, a competitive scheme with real-time rank feedback, and an incentive scheme that combines an individual piece rate with performance-contingent donations to charity outperforms a universal roll-out of the loss-framed bonus scheme that achieved the highest average performance in a first experiment even in a setting with little easily detectable heterogeneity (DellaVigna and Pope, forthcoming).

Yet, our results also highlight limitations of the approach. The algorithmic assignment performs better on a sample of individuals who provide reliable answers to elicited personality traits. Researchers and practitioners intending to use algorithmic assignment in different applications should therefore make sure that the quality of survey responses is high across individuals.

Turning to our specific application, future research should investigate the difference between workers own selection/sorting into different incentive schemes and the algorithmic assignment in more detail. It is conceivable that the preferences of workers for different incentives differ from what is best to increase their performance. Yet, it remains of interest whether sorting is a suitable alternative for workers that are very different from the original training sample.

Our results have several implications for the design of incentive schemes. Organizations can use individual worker characteristics to assign incentive schemes that in turn increase workers' performance. Moreover, 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.

A potential challenge might be the elicitation of the relevant characteristics to properly assign the best incentive scheme. Given that increasing amounts of data are available, this challenge is probably less severe than it was a few years ago. However, 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.

References

Andersson, Ola, Håkan J Holm, Jean-Robert Tyran, and Erik Wengström. 2016. "Deciding for Others Reduces Loss Aversion." *Management Science*, 62(1): 29–36.

Armantier, Olivier, and Amadou Boly. 2015. "Framing of Incentives and Effort Provision." *International Economic Review*, 56(3): 917–938.

Ashraf, Nava, Oriana Bandiera, and B Kelsey Jack. 2014. "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery." *Journal of Public Economics*, 120: 1–17.

Athey, Susan, and Guido W Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences*, 113(27): 7353–7360.

Athey, Susan, and Guido W Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives*, 31(2): 3–32.

Athey, Susan, and Guido W Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics*, 11(1): 685–725.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *The Quarterly Journal of Economics*, 120(3): 917–962.

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

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

Banker, Rajiv D, Seok-Young Lee, Gordon Potter, and Dhinu 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, Iwan. 2012. "Rank incentives: Evidence from a Randomized Workplace Experiment." https://repository.upenn.edu/bepp_papers/75, Accessed: 2022-03-01.

Battocchi, Keith, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. "EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation." https://github.com/microsoft/EconML, Version 0.x.

Benet-Martínez, Verónica, and Oliver 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.

Blanes i Vidal, Jordi, and Mareike Nossol. 2011. "Tournaments Without Prizes: Evidence from Personnel Records." *Management Science*, 57(10): 1721–1736.

Breiman, Leo. 2001. "Random Forests." *Machine learning*, 45(1): 5–32.

Butschek, Sebastian, Roberto González Amor, Patrick Kampkötter, and Dirk Sliwka. 2021. "Motivating Gig Workers–Evidence from a Field Experiment." *Labour Economics*, 102105.

Cadsby, C Bram, Fei Song, and Francis Tapon. 2007. "Sorting and Incentive Effects of Pay for Performance: An Experimental Investigation." *Academy of Management Journal*, 50(2): 387–405.

Caria, Stefano, Grant Gordon, Maximilian Kasy, Simon Quinn, Soha Shami, and Alex Teytelboym. 2020. "An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan." CESifo Working Paper No. 8535.

Carpenter, Jeffrey, and Erick Gong. 2016. "Motivating Agents: How Much Does the Mission Matter?" *Journal of Labor Economics*, 34(1): 211–236.

Casas-Arce, Pablo, and F Asis Martinez-Jerez. 2009. "Relative Performance Compensation, Contests, and Dynamic Incentives." *Management Science*, 55(8): 1306–1320.

Chen, Daniel L, Martin Schonger, and Chris Wickens. 2016. "oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance*, 9: 88–97.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018*a*. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal*, 21(1): C1–C68.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018*b*. "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." National Bureau of Economic Research Working Paper 24678.

Czibor, Eszter, Danny Hsu, David Jimenez-Gomez, Susanne Neckermann, and Burcu Subasi. 2022. "Loss-Framed Incentives and Employee (Mis-) Behavior." *Management Science*, 0(0).

Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke. 2013. "Tournament Incentives in the Field: Gender Differences in the Workplace." *Journal of Labor Economics*, 31(2): 305–326.

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *The Review of Economic Studies*, 85(2): 1029–1069.

DellaVigna, Stefano, and Devin Pope. forthcoming. "Stability of Experimental Results: Forecasts and Evidence." *American Economic Journal: Microeconomics*.

DellaVigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao. 2022. "Estimating Social Preferences and Gift Exchange at Work." *American Economic Review*, 112(3): 1038–1074.

De Quidt, Jonathan, Francesco Fallucchi, Felix Kölle, Daniele Nosenzo, and Simone Quercia. 2017. "Bonus versus Penalty: How Robust are the Effects of Contract Framing?" *Journal of the Economic Science Association*, 3(2): 174–182.

Dohmen, Thomas, and Armin Falk. 2011. "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender." *American Economic Review*, 101(2): 556–90.

Donato, Katherine, Grant Miller, Manoj Mohanan, Yulya Truskinovsky, and Marcos 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, Michalis, and Mary L Rigdon. 2022. "Gender Differences in Competitiveness: The Role of Social Incentives." CESifo Working Paper No. 9518.

Dubé, Jean-Pierre, and Sanjog Misra. forthcoming. "Personalized Pricing and Consumer Welfare." *Journal of Political Economy*.

Englmaier, Florian, and Stephen Leider. 2020. "Managerial Payoff and Gift-Exchange in the Field." *Review of Industrial Organization*, 56(2): 259–280.

Eyring, Henry, 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, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *The Quarterly Journal of Economics*, 133(4): 1645–1692.

Falk, Armin, Anke Becker, Thomas J Dohmen, David Huffman, and Uwe Sunde. 2016. "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences." IZA Discussion Paper No. 9674.

Fallucchi, Francesco, Daniele Nosenzo, and Ernesto Reuben. 2020. "Measuring Preferences for Competition with Experimentally-Validated Survey Questions." *Journal of Economic Behavior & Organization*, 178: 402–423.

Farrell, Anne M, Jonathan H Grenier, and Justin 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, Max H, Tengyuan Liang, and Sanjog Misra. 2021. "Deep Neural Networks for Estimation and Inference." *Econometrica*, 89(1): 181–213.

Ferraro, Paul J, and J Dustin Tracy. 2021. "A Reassessment of the Potential for Loss-Framed Incentive Contracts to Increase Productivity: A Meta-Analysis and a Real-Effort Experiment." ESI Working Paper 21-20. https://digitalcommons.chapman.edu/esi_working_papers/357/.

Foster, Jared C, Jeremy M G Taylor, and Stephen J Ruberg. 2011. "Subgroup Identification from Randomized Clinical Trial Data." *Statistics in Medicine*, 30(24): 2867–2880.

Friebel, Guido, Matthias Heinz, Miriam Krueger, and Nikolay Zubanov. 2017. "Team Incentives and Performance: Evidence from a Retail Chain." *American Economic Review*, 107(8): 2168–2203.

Gächter, Simon, Eric J Johnson, and Andreas Herrmann. 2021. "Individual-Level Loss Aversion in Riskless and Risky Choices." *Theory and Decision*, 1–26.

Gibbons, Frederick X, and Bram 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, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences." *The Quarterly Journal of Economics*, 118(3): 1049–1074.

Grolleau, Gilles, Martin G Kocher, and Angela Sutan. 2016. "Cheating and Loss Aversion: Do People Cheat More to Avoid a Loss?" *Management Science*, 62(12): 3428–3438.

Hannan, R. Lynn, Vicky B. Hoffman, and Donald V. Moser. 2005. "Bonus versus Penalty: Does Contract Frame Affect Employee Effort?" In *Experimental Business Research*. , ed. Amnon Rapoport and Rami Zwick, 151–169. Boston, MA:Springer US.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2, Springer.

Hirano, Keisuke, and Jack R Porter. 2009. "Asymptotics for Statistical Treatment Rules." *Econometrica*, 77(5): 1683–1701.

Hitsch, Günter J, and Sanjog Misra. 2018. "Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation." Available at SSRN 3111957.

Horton, John J, David G Rand, and Richard J Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics*, 14(3): 399–425.

Hossain, Tanjim, and John A List. 2012. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Science*, 58(12): 2151–2167.

Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *The Annals of Applied Statistics*, 7(1): 443–470.

Imas, Alex. 2014. "Working for the "Warm Glow": On the Benefits and Limits of Prosocial Incentives." *Journal of Public Economics*, 114: 14–18.

Imas, Alex, Sally Sadoff, and Anya Samek. 2017. "Do People Anticipate Loss Aversion?" *Management Science*, 63(5): 1271–1284.

John, Oliver P, Eileen M Donahue, and Robert L Kentle. 1991. "Big Five Inventory–Versions 4a and 54." Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, Oliver P, Laura P Naumann, and Christopher J Soto. 2008. "Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues." In *Handbook of Personality: Theory and Research.* , ed. Oliver P John, Richard W Robins and Lawrence A Pervin, 114–158. The Guilford Press.

Katz, Lawrence F., and Alan B. Krueger. 2019. "The Rise and Nature of Alternative Work Arrangements in the United States, 1995–2015." *ILR Review*, 72(2): 382–416.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica*, 86(2): 591–616.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions*." *The Quarterly Journal of Economics*, 133(1): 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review*, 105(5): 491–95.

Larkin, Ian, and Stephen Leider. 2012. "Incentive Schemes, Sorting, and Behavioral Biases of Employees: Experimental Evidence." *American Economic Journal: Microeconomics*, 4(2): 184–214.

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review*, 90(5): 1346–1361.

Lazear, Edward P. 2018. "Compensation and Incentives in the Workplace." *Journal of Economic Perspectives*, 32(3): 195–214.

Levitt, Steven D, John A List, Susanne Neckermann, and Sally Sadoff. 2016. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." *American Economic Journal: Economic Policy*, 8(4): 183–219.

Manthei, Kathrin, Dirk Sliwka, and Timo Vogelsang. 2021. "Performance Pay and Prior Learning—Evidence from a Retail Chain." *Management Science*, 67(11): 6998–7022.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete too much?" *The Quarterly Journal of Economics*, 122(3): 1067–1101.

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.

Rammstedt, Beatrice, and Oliver 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, Holger A. 2014. "The Disposition Effect and Loss Aversion: Do Gender Differences Matter?" *Economics Letters*, 123(1): 33–36.

Snowberg, Erik, and Leeat Yariv. 2021. "Testing the Waters: Behavior across Participant Pools." *American Economic Review*, 111(2): 687–719.

Sprinkle, Geoffrey B, and Michael G Williamson. 2006. "Experimental Research in Managerial Accounting." *Handbooks of Management Accounting Research*, 1: 415–444.

Tonin, Mirco, and Michael Vlassopoulos. 2010. "Disentangling the Sources of Pro-Socially Motivated Effort: A Field Experiment." *Journal of Public Economics*, 94(11-12): 1086–1092.

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

Van der Stede, Wim A, Anne Wu, and Steve Yu-Ching Wu. 2020. "An Empirical Analysis of Employee Responses to Bonuses and Penalties." *The Accounting Review*, 95(6): 395–412.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

A Appendix A: Experimental Design

Survey Details

- Demographics (age, gender, education level)
- Big-5 (Benet-Martínez and John 1998, John, Donahue and Kentle 1991, John, Naumann and Soto 2008, Rammstedt and John 2007)
- Risk preferences (Falk et al. 2016, 2018)
- Loss aversion (Gächter, Johnson and Herrmann 2021)
- Competitiveness (Fallucchi, Nosenzo and Reuben 2020)
- Social comparison (Gibbons and Buunk 1999)
- Altruism (Falk et al. 2016, 2018)
- Positive reciprocity (Falk et al. 2016, 2018)

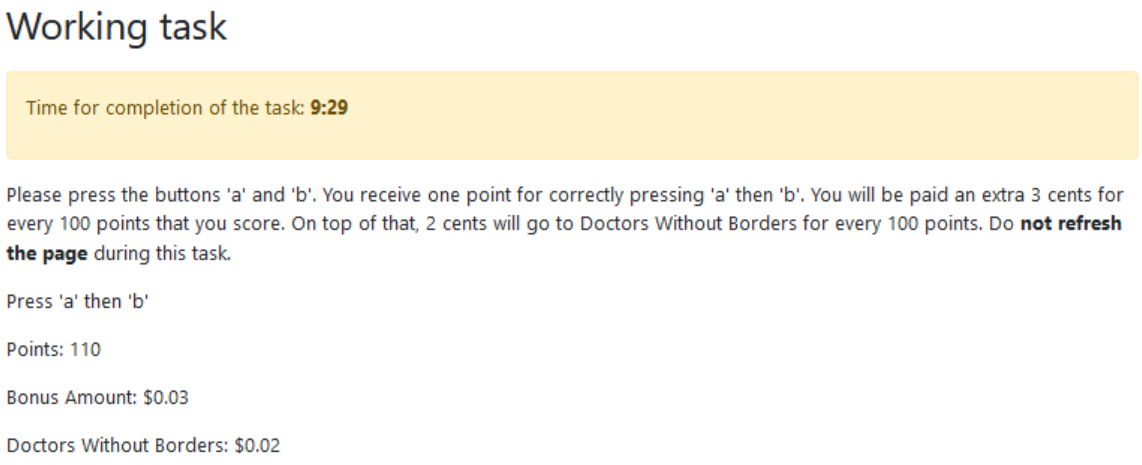


Figure 4: Screenshot of the Working Stage (*Social PfP* Treatment)

B Appendix B: Results

Table 9: 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 10: Algorithm Comparison

	Residualized Performance			
	Bonus Loss	RTR Feedback	Social PfP	Overall
	(1)	(2)	(3)	(4)
Indirect Random Forest	126.9	133.6	153.7	132.8
(Share of Obs.)	(46.8)	(41.6)	(11.6)	(100.0)
Causal Forest	129.7	122.0	117.5	126.0
(Share of Obs.)	(55.6)	(39.1)	(5.3)	(100.0)
Doubly Robust	130.4	133.2	97.8	128.3
(Share of Obs.)	(49.4)	(40.7)	(9.9)	(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 11: Results of Experiment 1

	$log(Performance)_i$				
	(1)	(2)	(3)	(4)	(5)
PfP_i	0.375***	0.377***	-0.031	-0.024	0.043
	(0.042)	(0.042)	(0.034)	(0.024)	(0.037)
$Bonus\ Gain_i$	0.359***	0.360***	-0.048	-0.040	0.027
	(0.050)	(0.050)	(0.044)	(0.045)	(0.050)
$Gift\ and\ Goal_i$	0.210***	0.209***	-0.198***	-0.191***	-0.124**
	(0.052)	(0.052)	(0.047)	(0.044)	(0.047)
$Bonus\ Loss_i$	0.403***	0.408***		0.008	0.075*
	(0.047)	(0.047)		(0.035)	(0.039)
$RTR\ Feedback_i$	0.394***	0.400***	-0.008		0.067*
	(0.042)	(0.041)	(0.035)		(0.037)
$Social\ PfP_i$	0.330***	0.333***	-0.075*	-0.067*	
	(0.051)	(0.051)	(0.039)	(0.037)	
$Control_i$			-0.408***	-0.400***	-0.333***
			(0.047)	(0.041)	(0.051)
Dem. Controls	No	Yes	Yes	Yes	Yes
Observations	6,065	6,065	6,065	6,065	6,065
Adjusted R-squared	0.125	0.128	0.128	0.128	0.128

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 columns(1) and (2), we use the control group as reference group, thus reporting the treatment effects for the different incentive schemes in comparison to the control group. 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 (2) to (5), we further include demographics, i.e. age, gender dummmies and education dummies as controls. In column (3) to (5), we use the *Bonus Loss*, the *Real-time Rank Feedback* and the *Social PfP* treatment as reference group, respectively. Standard errors are clustered on batch level, and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 12: Robustness Check: Effect on Performance

	$Performance_i$			
	(1)	(2)	(3)	(4)
$Algorithm_i$	343.073***	342.424***	31.325**	31.466**
	(53.865)	(54.251)	(13.799)	(13.862)
$Best\ ATE_i$	311.671***	310.848***		
	(53.617)	(53.857)		
P-value Best ATE=Algorithm	0.027	0.027		
Reference Group	Control	Control	Best ATE	Best ATE
Dem. Controls	No	Yes	No	Yes
Observations	6,377	6,377	6,147	6,147
Adjusted R-squared	0.178	0.180	0.173	0.176

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 (3) and (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 (2) and (4), we further include age, gender dummies as well as education dummies as controls. Standard errors are clustered on batch level, and reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table 13: Sample Differences between Retakers and New Hires

	Retakers		*New Hires*		
	Mean	S.D.	Mean	S.D.	p-value
Ability Proxy	46.054	20.435	41.568	22.573	0.000
Age	40.711	12.250	37.740	11.640	0.000
Female	0.451	0.498	0.513	0.500	0.000
Some College	0.141	0.348	0.170	0.375	0.004
Bachelor's Degree or more	0.771	0.420	0.715	0.452	0.000
Consistency (Z-scored)	0.099	1.021	-0.048	0.986	0.000
Observations	2096		4282		6378

Note: In this table, we report the summary statistics of the 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 consistency as well as the p-values of a test of proportions for the binary variables, 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. The proxy 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).

Table 14: Consistency Comparison between Retakers and Other Workers

	$Consistency_i$	
	Experiment 2	Experiment 1
	(1)	(2)
$Retaker\ (Exp2)_i$	0.255***	0.212***
	(0.035)	(0.026)
Dem. Controls	Yes	Yes
Observations	6,377	6,065
Adjusted R-squared	0.173	0.217

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 1, i.e. the reference group for the retakers (exp2) are the workers taking part in experiment 1 only. In column (2), we restrict the samples to experiment 2, i.e. the reference group for the retakers (exp2) are the new hires. We include batch fixed effects, an ability proxy as well as demographics, i.e. age, gender dummies and education dummies, 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.

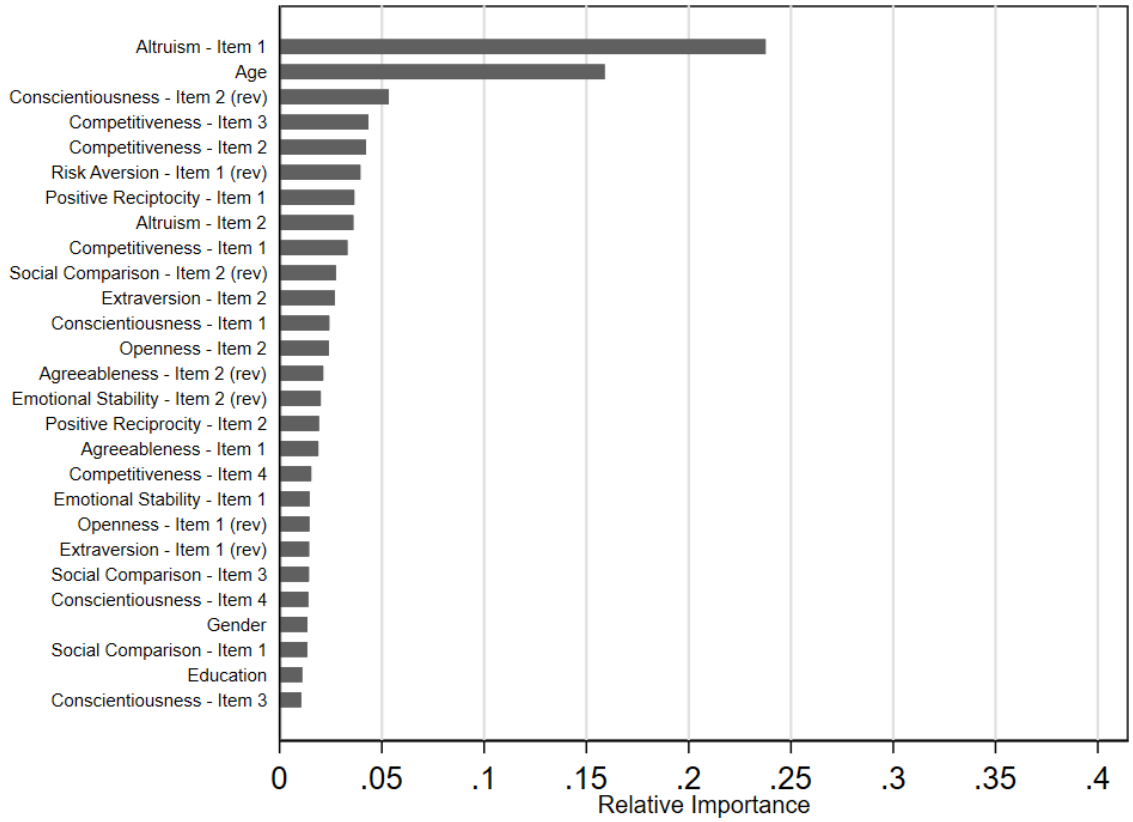


Figure 5: Feature Importances - Bonus Loss

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, for example, chapter 10 of Hastie, Tibshirani and Friedman (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.

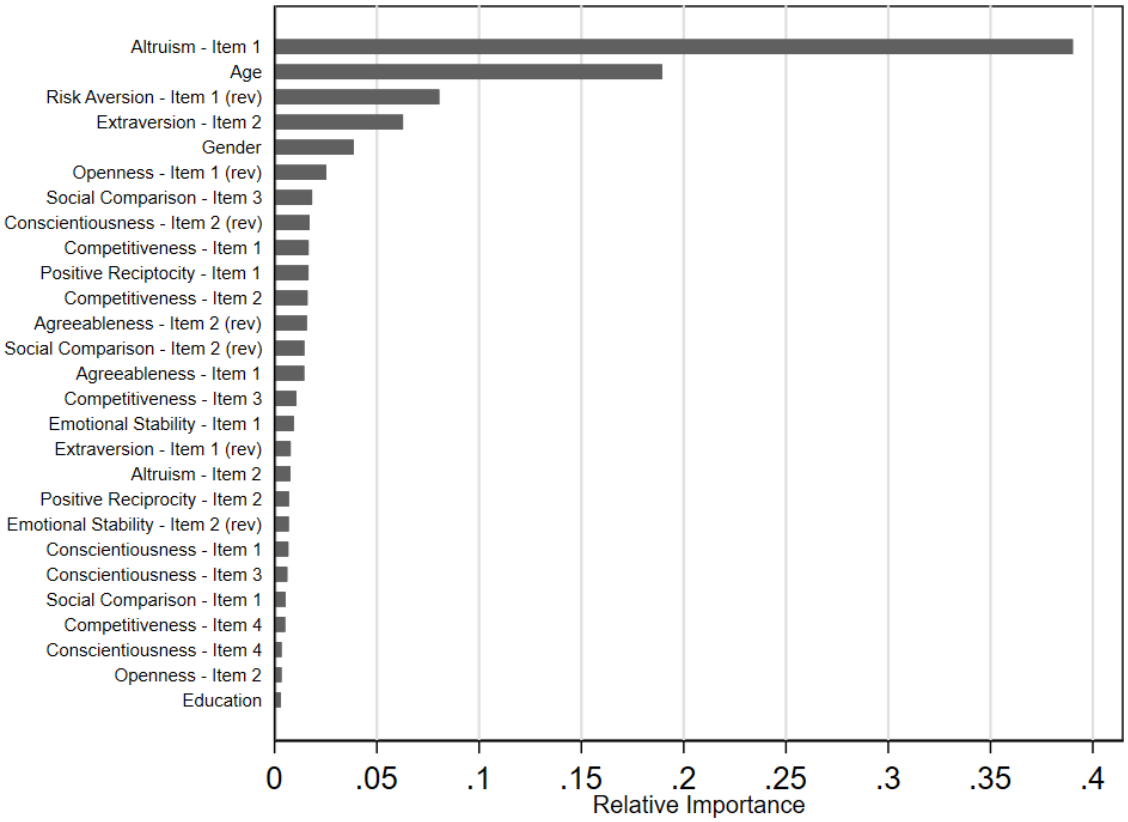


Figure 6: Feature Importances - RTR Feedback

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, for example, chapter 10 of Hastie, Tibshirani and Friedman (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.

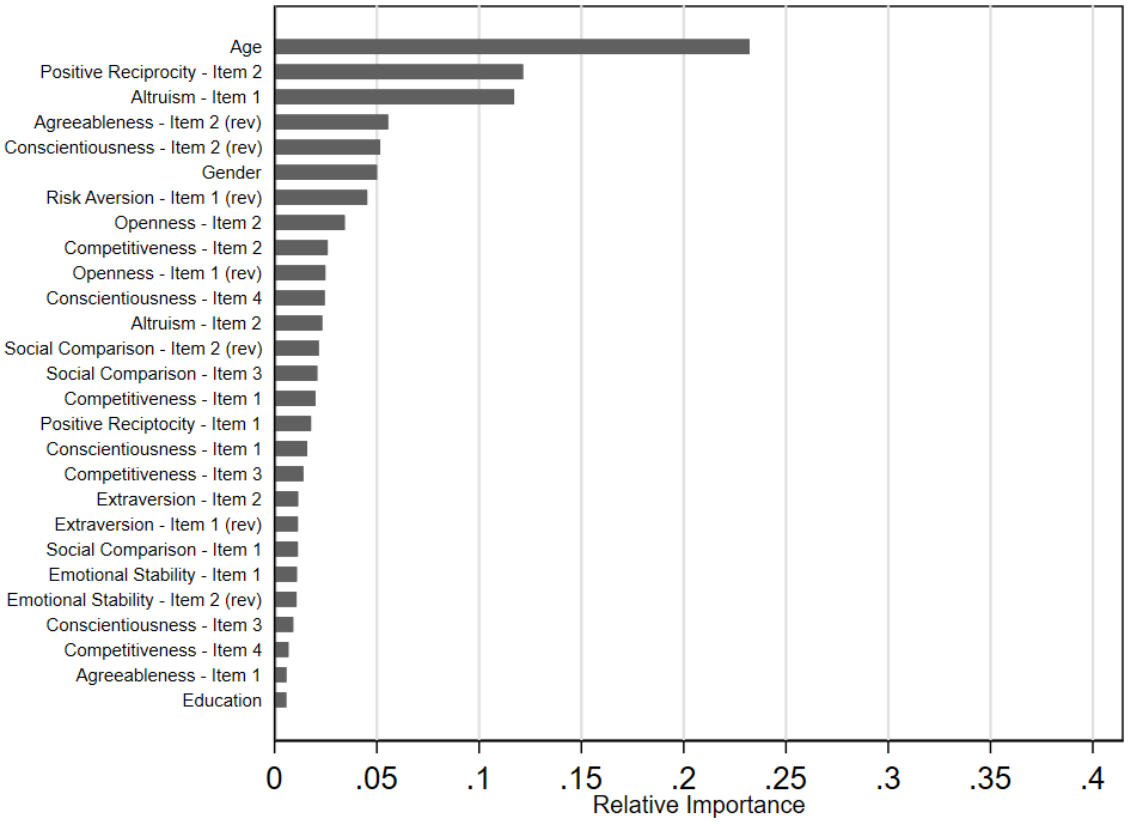
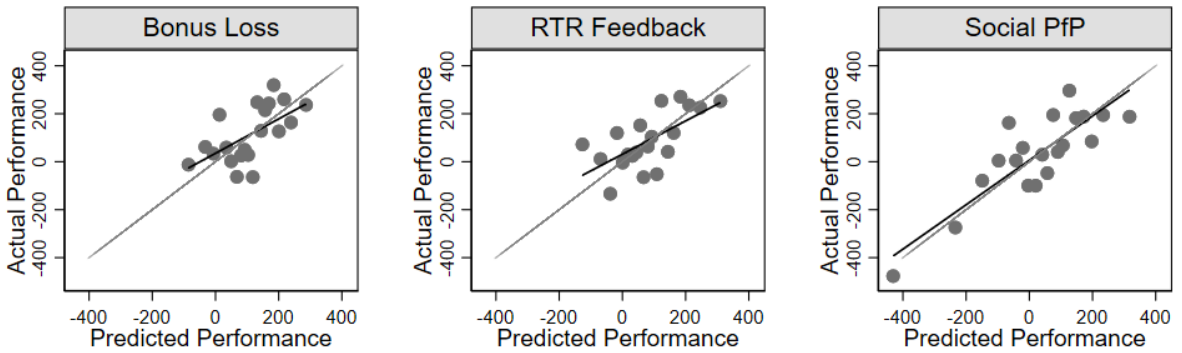
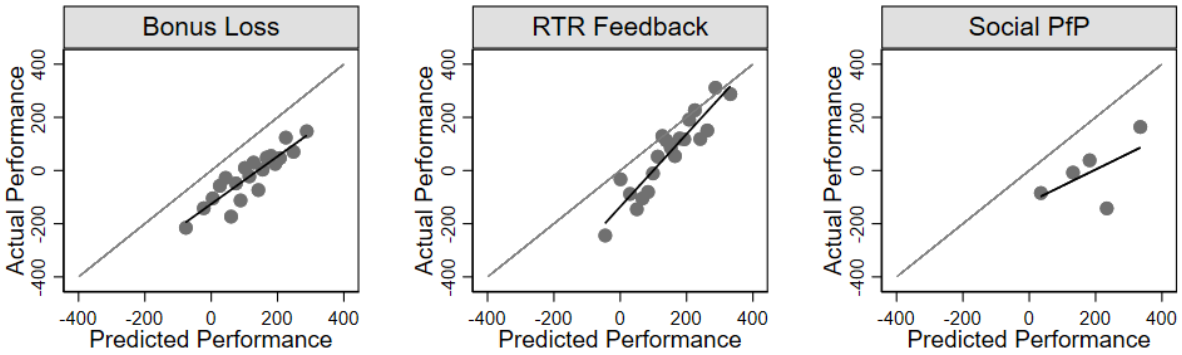


Figure 7: Feature Importances - Social PfP

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, for example, chapter 10 of Hastie, Tibshirani and Friedman (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.



(a) Experiment 1



(b) Experiment 2

Figure 8: 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.

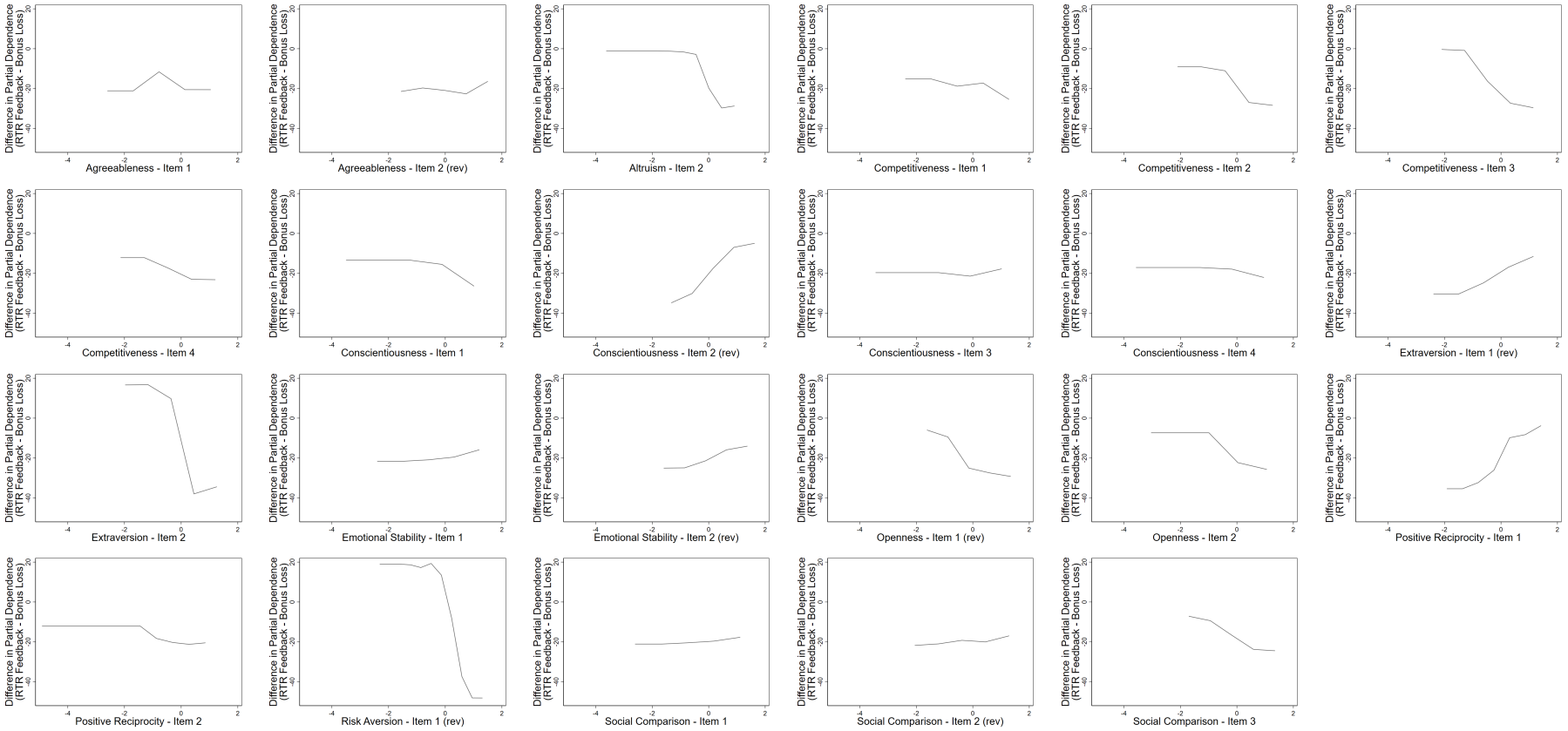


Figure 9: Partial Dependence Comparisons (RTR Feedback - Bonus Loss)

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.

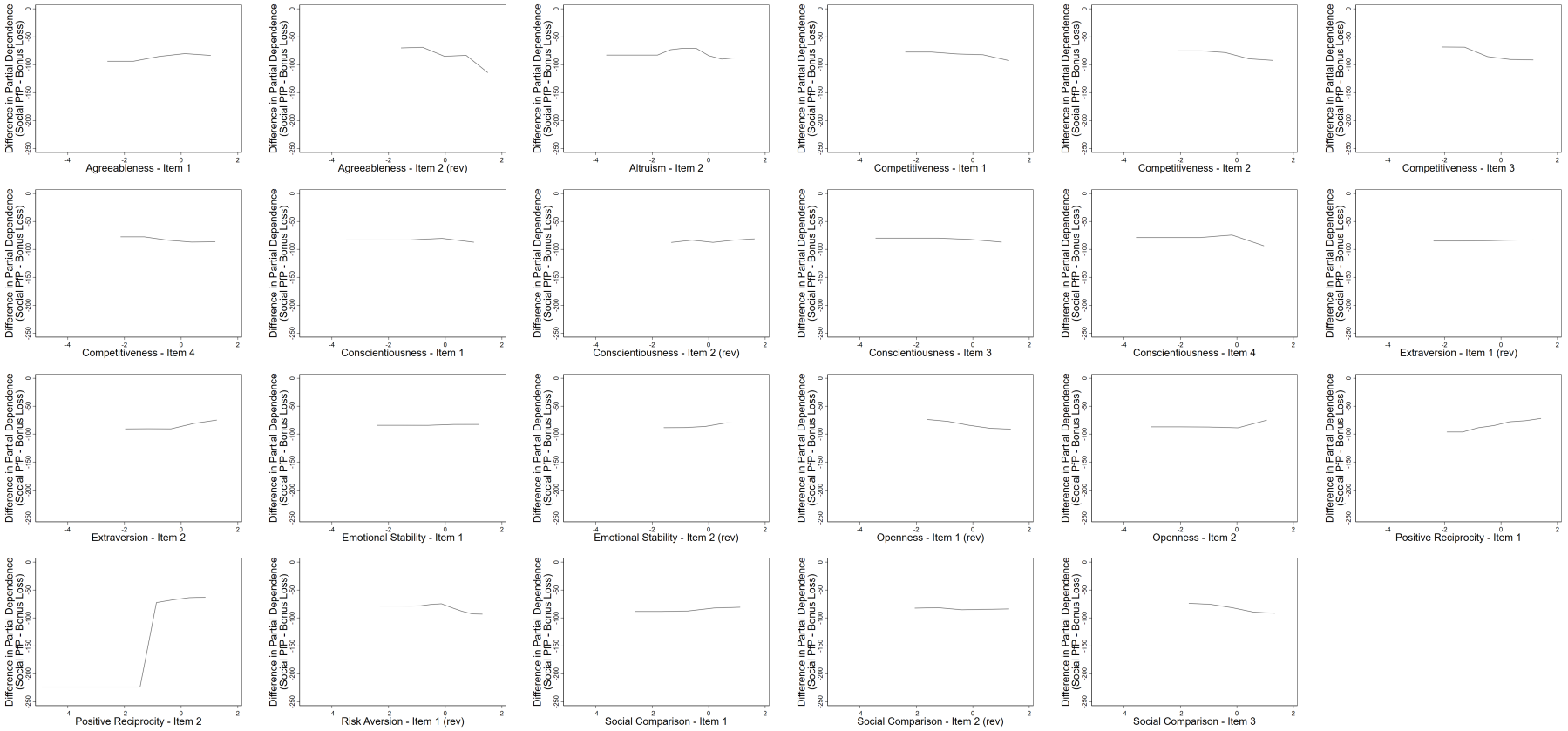


Figure 10: Partial Dependence Comparisons (Social PfP - Bonus Loss)

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.

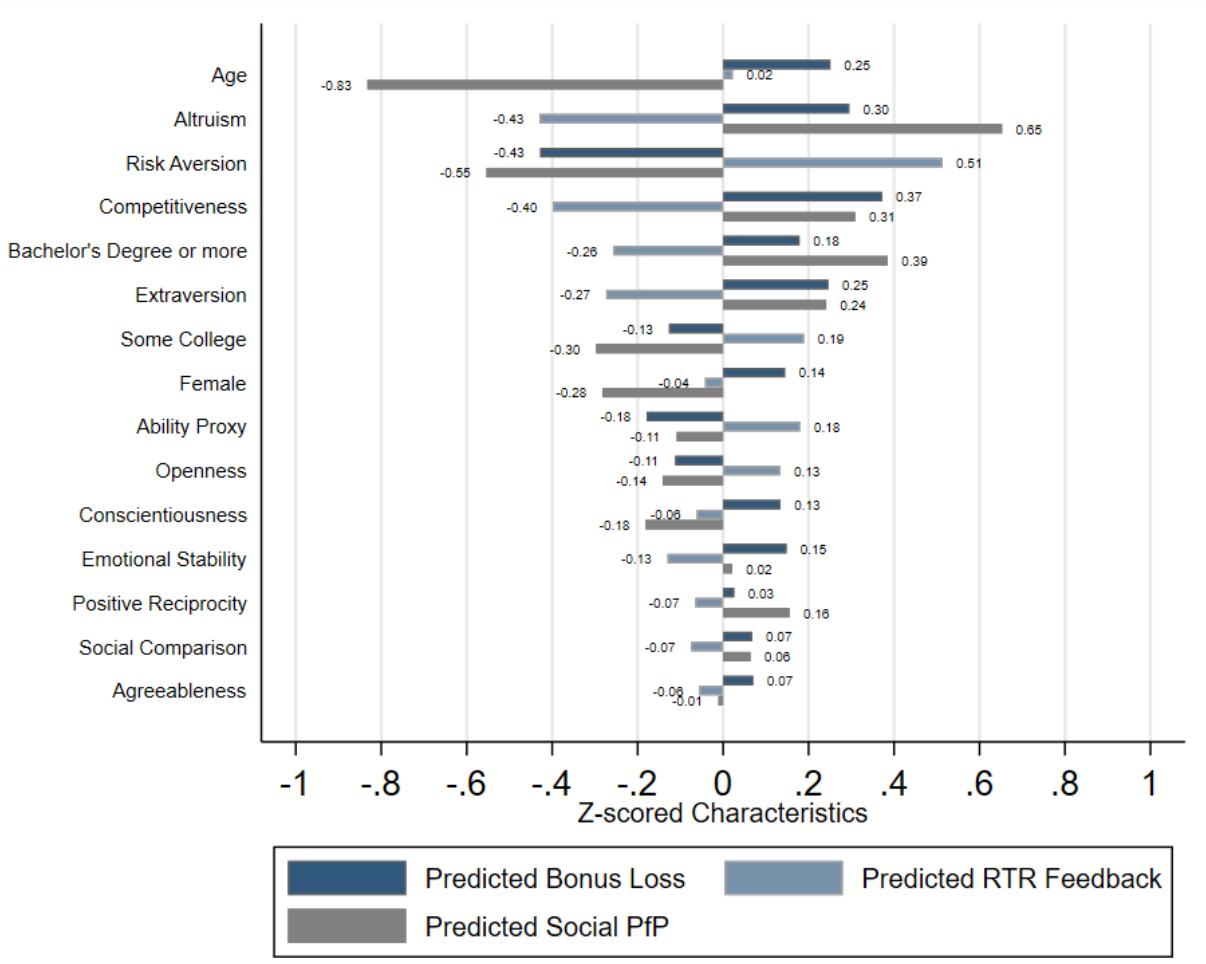


Figure 11: Group Characteristics

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.