NBER WORKING PAPER SERIES

THE IMPACT OF JOINT VERSUS SEPARATE PREDICTION MODE ON FORECASTING ACCURACY

Alex Imas
Minah H. Jung
Silvia Saccardo
Joachim Vosgerau

Working Paper 30611
http://www.nber.org/papers/w30611

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2022

We are grateful for helpful comments from Stefano Dellavigna, Chris Hsee, and Devin Pope. This research received no external funding. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Impact of Joint versus Separate Prediction Mode on Forecasting Accuracy
Alex Imas, Minah H. Jung, Silvia Saccardo, and Joachim Vosgerau
NBER Working Paper No. 30611
October 2022
JEL No. D0,D9,D90

ABSTRACT

Forecasters predicting how people change their behavior in response to a treatment or intervention often consider a set of alternatives. In contrast, those who are treated are typically exposed to only one of the treatment alternatives. For example, managers selecting a wage schedule consider a set of alternative wages while employees are hired at a given rate. We show that forecasts made in joint-prediction mode—which considers a set of alternatives—generate predictions that expect substantially larger behavioral responses than those made in separate-prediction mode—which considers the response to only one treatment realization in isolation. Results show the latter to be more accurate in matching people's actual responses to interventions and treatment changes. We present applications to managerial decision-making and forecasting of scientific results.

Alex Imas
Booth School of Business
University of Chicago
5897 S. Woodlawn Avenue
Chicago, IL 60637
and NBER
alex.imas@chicagobooth.edu

Minah H. Jung
New York University
minah.jung@stern.nyu.edu

Silvia Saccardo
Department of Social and Decision Sciences
Carnegie Mellon University
5000 Forbes Avenue
BP 208
Pittsburgh, PA 15213
ssaccard@andrew.cmu.edu

Joachim Vosgerau
Università Bocconi
Via Röntgen 1
20136 Milano, Italia
joachim.vosgerau@unibocconi.it

1 Introduction

Prediction is a central feature of decision-making. Choice under uncertainty involves predicting outcomes given one's beliefs about their likelihoods and deciding accordingly. Such predictions are assumed to be accurate on average given the information at hand in the majority of economic models (Mishkin, Ungerleider, and Macko, 1983). An important subset of choice under uncertainty involves a principal making forecasts about how changes in a variable impact agents' action. For example, a manager deciding whether or not to increase wages must weigh each option's cost against the predicted effect on workers' performance; a policy maker implementing a new tax must evaluate the increased revenues of each rate against potential inefficiencies-both stemming from the forecasted consumer response; a scientist designing an experiment must forecast the effect of treatment variables on participants' actions. In many important settings, the principal makes such forecasts in a different evaluation mode than the agent: while the principal predicts an agent's response across a set of potential treatments (e.g., different wages), the agent makes a decision in response to one realization (e.g. exerting effort in response to one wage).

Why would this distinction matter? From the standpoint of standard economic theory, it should not. The agent makes an effort choice based on how the variable of interest interacts with her preferences; the principal makes forecasts based on the correct mental model of the agent's preferences. Research from psychology, however, suggests that the two evaluation modes—one that evaluates responses to variables jointly and the other to each variable in isolation—may lead to a systematic wedge between forecasts and behavior (Hsee, Loewenstein, Blount, and Bazerman, 1999).

While agents may respond to specific variable realizations, the principal naturally compares different realizations to one another. The set of alternatives considered by the principal may be systematically different than the one considered by the agent, who would generate this set through spontaneous memory recall. If the agent recalls a more compact set, or does not engage in counterfactual thinking at all, then the principal's forecast may be systematically biased. Notably, research on the contrast effect (Pepitone and DiNubile, 1976; Hartzmark and Shue, 2018) and on the effects of anchoring on valuations (Ariely, Loewenstein, and Prelec, 2003) suggests that joint prediction mode may lead principals to forecast larger responses to variable changes than what is actually observed. This leads to the question of how forecasts in a joint prediction mode would compare to those in a separate prediction mode, in which principals—like the agents—evaluate each level of the variable in isolation. Principals in a separate prediction mode face similar information environments as the agents, and would likely recall a similar set of alternatives when predicting behavioral responses. Adopting a separate prediction mode may thus align the principal's mental model more closely with that of the agent, which would generate more accurate forecasts of behavior change.

We examine this conjecture in three experimental studies[1]. In the first Experiment (1A), participants from an online crowdsourcing platform were assigned to the role of forecasters or workers. Workers performed a real-effort task at one of two different piece rates without being informed about the possibility of the other. Forecasters were given information about the task and incentivized to predict worker performance in one of three treatments. In the Joint-Prediction treatment, the forecasters reported their belief about worker performance in response to each of the two wage levels. In the Separate-Prediction treatments, two separate groups of forecasters reported their beliefs about performance at one of the two wage levels; importantly, like the workers, forecasters in the Separate-Prediction treatments were not informed about the other wage level.

We find substantial differences between prediction modes. While forecasters in the Separate-Prediction treatments predicted modest differences in performance as the wage levels increased, averaging 15.0% of one standard deviation in effort, those in the Joint-Prediction treatments predicted much larger responses, averaging 68.4% of one standard deviation. Importantly, workers' actual performance differences, which averaged to 20.8% if one standard deviation in effort, were considerably closer to the forecasts of those in the Separate-Prediction treatments; the joint prediction mode led forecasters to exacerbate worker's responses to wage increases to an extent that was both statistically and economically significant.

The second Experiment (1B) sought to replicate these results using a different real-effort task and demonstrate how differences in prediction mode impact principals' decisions. We first replicated the results of the first study using three different wage levels and a different, widely used, real-effort task. We then recruited a separate group of participants to take on the role of managers hiring workers for a task. Managers were told that workers had performed the task at one of three wage levels, and that the workers did not know the existence of the other two. After reading this information, managers were asked to choose a wage level that would maximize their profit, where profit was

[1] All study instructions, materials, data, code, and preregistrations can be found at https://researchbox.org/829&PEER_REVIEW_passcode=ZKBEIL

a negative function of the wage paid and a positive function of the workers' actual average performance at that wage. Note that choosing a profit-maximizing wage involved correctly forecasting the workers' responses to wage increases. Consistent with the joint prediction mode exacerbating performance differences, managers overwhelmingly chose to pay higher wages than the profit-maximizing level.

Having shown the judgment and decision-making consequences of joint versus separate prediction modes, we proceeded to test our hypotheses in the context of forecasting scientific results. Recent efforts to improve the scientific process have included large-scale replications of prior studies paired with predictions from experts and non-experts on the success of these replications (Dreber, Pfeiffer, Almenberg, Isaksson, Wilson, Chen, Nosek, and Johannesson, 2015; Gordon, Viganola, Dreber, Johannesson, and Pfeiffer, 2021). Researchers have also argued that predictions of scientific findings may be useful more broadly as they help gauge the contribution of a specific result relative to people's prior beliefs, thereby circumventing hindsight bias; for example, a null result in light of a predicted effect would be surprising and more likely to be published (DellaVigna, Pope, and Vivalt, 2019). This has led to the development of multiple public prediction markets such as the Social Science Prediction Platform, where people can sign up to predict the outcomes of specific experiments that have not been run yet. A growing number of research studies have incorporated expert or non-expert predictions in their methodologies, often reporting prediction data that are directly in contrast with the corresponding behavioral data (Milkman, Gromet, Ho, Kay, Lee, Pandiloski, Park et al., 2021a; Bessone, Campante, Ferraz, and Souza, 2022). Critically, forecasts in these settings are almost exclusively elicited in joint prediction mode.

In one of the first papers to elicit predictions about novel scientific findings, DellaVigna and Pope (2018) asked academic experts to forecast the performance of online participants in response to a variety of incentive schemes. Experts were given information about performance in response to monetary incentives and made predictions on how "psychological incentives" such as loss aversion, low probability lottery payments, and prosocial incentives would affect effort levels [2]. All forecasts were made in joint prediction mode, with experts reporting their beliefs across all incentive schemes. To examine the impact of prediction mode on forecasts of scientific results, in Experiment 2 we recruited participants from an online crowdsourcing platform to perform the task

[2] Results on how the different incentive schemes impacted performance can be found in Dellavigna and Pope (2018)

used in Dellavigna and Pope (2018) under a subset of the incentives tested in the original experiment, and additional participants from the same platform to forecast their performance. Forecasters made predictions in either Joint or Separate-Prediction mode. We looked at the effects of both standard and prosocial incentives, where the latter involve a piece rate that is donated to charity as a function of the worker's effort (Imas, 2014). DellaVigna and Pope (2018) found that experts predicted larger performance differences from increasing prosocial incentives than what was actually observed.

Consistent with the findings from the first two studies, forecasters predicted larger treatment differences in the Joint-Prediction treatment than in the Separate-Prediction treatments. Forecasters in the Joint-Prediction treatment overstated the impact of prosocial incentives—replicating Dellavigna and Pope (2018) - and standard incentives relative to observed behavior. Forecasters in the Separate-Prediction treatment were much more accurate in predicting response differences to both types of incentives.

Together, our findings suggest that prediction mode significantly impacts the forecasting of scientific results and should therefore be taken into account when interpreting elicited beliefs. For example, recent work showing that behavioral scientists overestimate the impact of interventions may be driven by a discrepancy in prediction modes (Bowen, 2022).

Our findings contribute to the literature on joint versus separate evaluation (Hsee et al., 1999). This literature has documented systematic preference reversals when people evaluate options jointly or in isolation. For example, Bazerman, Loewenstein, and White (1992) show that people evaluating social allocations put more weight on relative payoffs between themselves and others when making decisions in isolation, but focus on their own payoffs when multiple allocations are evaluated jointly. Similarly, Blount and Bazerman (1996) found that people were more concerned about fairness in the workplace when making joint versus separate decisions. Hsee (1996) demonstrated preference reversals when evaluating several job candidates either as part of a set or in isolation, while Bohnet, van Geen, and Bazerman (2016) showed that group-based stereotypes are less likely to affect hiring decisions when candidates are evaluated jointly versus separately. We add to this literature by demonstrating the impact of evaluation mode on predictions. Namely, while prior work has focused on choice, we show that joint prediction mode leads people to forecast larger causal effects of variables (e.g. how increasing wages impact effort) than separate prediction mode.

This work also relates to research documenting a discrepancy in between-subject versus within-subject experimental designs. For example, Fox and Tversky (1995) find evidence of ambiguity aversion over gambles in a within-subject design, but not when participants made choices over the same gambles in isolation in a between-subject design. In the context of willingness to pay elicitation, Frederick and Fischhoff (1998) find substantially larger effect sizes in within as opposed to between-subject design (see also Charness, Gneezy, and Kuhn, 2012).

Finally, our research is related to the work on mental representation and counterfactual thinking (Medvec, Madey, and Gilovich, 1995; Roese, 1997). While forecasters in the joint prediction mode make predictions over a set of treatment possibilities, those in the separate prediction mode explicitly consider only one realization; importantly, the latter faces a similar informational environment as the agents. Our findings suggest that both agents and forecasters in separate prediction mode do not spontaneously recall the same alternative realizations as those in the joint prediction mode—if alternative realizations are recalled at all. If a similar set of alternatives was recalled and considered, then forecasts would not differ by prediction mode. This points to differential recall as a potential wedge between the two prediction modes (Gennaioli and Shleifer, 2010; Bordalo, Gennaioli, Ma, and Shleifer, 2020) and a potential methodology for improving accuracy. Namely, the difference in accuracy between prediction modes highlights a misspecification in the forecasters' mental models of the agent's decision problem (Bohren and Hauser, 2021), as forecasters in joint prediction mode do not seem to account for the agents' failure to consider a similar set of alternative treatment realizations when making their decisions. Our results suggest that aligning the forecasters' and agents' mental models by ensuring that the former is provided with the same set of treatment alternatives as the latter can improve prediction accuracy.

2 Experiment 1A and 1B

In our first two experiments, we test whether prediction mode—jointly predicting effort levels across wage levels or separately predicting effort levels within each wage level—affects prediction accuracy and realized profits. To this end, in each of the two experiments, we hired workers to complete a real-effort task in one of several randomly-assigned wage conditions.

We then separately recruited groups of forecasters to predict the observed average effort levels in each wage condition. The first group predicted average effort levels in the Joint Prediction (JP) treatment across all wage levels. The second group predicted average effort levels in the Separate Prediction (SP) treatment, such that each individual made a prediction only for one wage level. To examine the effect of prediction mode on behavior, the second study recruited a third group of participants to act as managers. Each manager chose to hire a worker who would be paid one of the wage levels with the intention of maximizing profits. The experimental instructions for these two experiments can be found in Online Appendices A and B.

We predicted that forecasters in the JP treatment would predict effort to be more wage-sensitive than forecasters in the SP treatments. As a consequence, managers (who made their decisions in the joint prediction mode) would choose higher than optimal wages given empirical estimates of the profit function.

2.1 Experiment 1A

Procedure: We recruited a total of 2,337 participants on Prolific. Of these, 1141 participants worked on a real-effort task and 1,196 forecasted the effort exerted in that task. We recruited and pre-registered the two samples separately[3].

The real-effort task consisted of collecting images for an online database (see Schwartz, Keenan, Imas, and Gneezy 2021 for previous use of this task). All workers received a base payment of $0.12 and were then randomly assigned to one of the two wage conditions, receiving additional bonus payments of $0.02 or $0.10 per every unique URL link to online images of exotic flowers they provided. Workers were given 1 hour to provide up to 30 URL links and could decide to end the search task at any time and complete the study. In order to provide the image links, we asked workers to type "exotic flower" into Google images and copy paste the Google image link that they would like to add to the database. To ensure that workers paid attention to the instructions, we included two attention checks asking workers how much time they had to complete the task.

Forecasters were randomly assigned to one of three treatments: a JP treatment and two SP treatments. All forecasters read a description of the flower image collection task and practiced the task themselves by providing one URL link of an image of exotic flowers. We then explained to forecasters that in a previous study "Participants were a mix of those who did and didn't have experience in completing tasks on the platform. The Prolific participants who completed the study were all from the United States. Among them, 62% were female with an average age of 35 years (50% of participants were between 18-31 years old, and 50% were over 32 years old)." Furthermore, forecasters were informed

[3] The real-effort experiment had been conducted as part of a different research project that was never used. Please see Appendix A for details.

that the number of URL links provided by workers ranged from 0 to 30. Forecasters in the JP treatment were then asked to predict the average number of URL links provided across the two wage levels ($0.02 per link and $0.10 per link); forecasters in the SP treatments were asked to either predict the average number of sliders completed by workers who received $0.02 per link or to predict the average number of sliders completed by workers who received $0.10 per link. All forecasters were informed that five randomly chosen forecasters would receive an additional payment based on the accuracy of their forecasts equal to ($100- mean square error/2,000) in additional to their base payment of $0.55.

Results: For workers, we excluded 67 responses of those who did the study twice and 163 responses of those those who did not correctly answer the two comprehension checks, leaving 911 responses for analyses. For forecasters, we excluded responses from 2 individuals who failed a comprehension check question three times in a row, leaving 1,194 responses for analyses.

Workers responded to increasing wages by providing significantly more flower links at $0.10 per link than at $0.02 per link (Difference in Actual Effort: $M = 2.571, t(909) = 3.142, p = 0.002, Cohen's\ d = 0.208)$[4].

How did forecasters' predictions compare to the workers' performance? Forecasters in the JP treatment predicted that workers would have a strong response to increases in wages (Difference in Predicted Effort: $M_{diff\_JP} = 8.442, t(794) = 16.434, p < 0.001, Cohen's\ d = 1.165$). Forecasters in the SP treatment also predicted that workers would respond to increases in wages (Difference in Predicted Effort: $M_{diff\_SP} = 1.853, t(794) = 3.494, p < 0.001, Cohen'sd = 0.248$). However, as shown in Column 1 of Table 1, the predicted response was substantially more muted than in the JP treatment.

[4] For clarity of exposition, we report here the simple analyses of treatment differences. For specific effort levels (observed and predicted) and the preregistered analysis of prediction accuracy (predicted – observed), see Appendix A.

Table 1. Forecasts in the JP and SP treatments in Experiment 1A and 1B

	Prediction	
	Experiment 1A	Experiment 1B
	(1)	(2)
Constant	14.21***	44.93***
	(0.39)	(1.39)
Separate Prediction	4.70***	10.57***
	(0.56)	(1.83)
$0.1 Incentive	8.44***	19.28***
	(0.51)	(1.90)
$0.1 Incentive × Separate Prediction	-6.59***	-14.64***
	(0.74)	(2.49)
$0.05 Incentive		8.86***
		(1.85)
$0.05 Incentive × Separate Prediction		-7.06***
		(2.49)
Observations	1592	1459

Notes: This table reports results from OLS regression, with robust standard errors in parentheses. The outcome measure is predicted effort in the forecasting experiments. Column 1 reports the results for Experiment 1A and Column 2 the results for Experiment 1B. Separate Prediction is an indicator variable coded as 1 when predictions were made in the Separate-Prediction treatment and 0 otherwise. $0.10 Incentive is an indicator variable coded as 1 when workers received the high ($0.10) wage in both experiments and 0 otherwise. $0.05 Incentive is an indicator variable coded as 1 when workers received the $0.05 in Experiment 1B, and 0 otherwise. Data from the JP treatments is treated as statistically independent to make it comparable to the forecasts in the SP treatments to observed (actual) effort.

Notably, as shown in Figure 1, forecasters in the SP treatments were also more accurate than those in the JP treatment. Predictions in the SP treatment did not differ from the actual effort responses to wages; in contrast, forecasters in the JP treatments substantially overestimated workers' responsiveness. Table A5 in Appendix A presents these results within a regression framework, showing that forecasts in the SP treatment were significantly more accurate in predicting responsiveness to different wage levels than those in the JP treatment. These results provide support for the notion that aligning forecasters' mental model with that of workers' leads to more accurate forecasts of treatment responses.

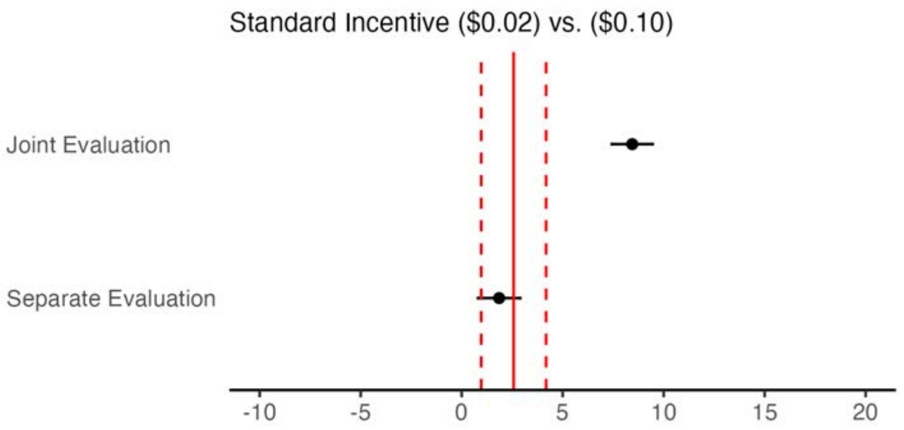


Figure 1. Forecasting Accuracy in Experiment 1A

Notes: The solid red line indicates the observed difference in the average number of URL links provided by workers in the $0.02 and $0.10 wage treatments (the dashed lines indicate the 95% confidence interval). The black dots indicate the forecasted differences in the average number of URL links provided by workers in the $0.02 and $0.10 wage conditions (the whiskers indicate 95% confidence intervals).

2.2 Experiment 1B

Procedure: Experiment 1b replicates Experiment 1a using a different real-effort task and further tests how differences in prediction mode impact forecasters' decisions. To this end, we recruited a total of 1,584 participants from Prolific. Of these, 301 participants were recruited in the role of workers and offered different wages for completing a real effort task that involved moving sliders across a computer screen. An additional 1,283 participants were recruited for the forecasting task. In the latter group, 1,165 forecasters predicted workers' exerted effort, and 118 participants were assigned to the role of managers who hired workers by choosing a wage. All three samples were recruited and pre-registered separately.

The workers in the slider task received a base payment of $0.55 and were randomly assigned to one of three wage conditions, receiving additional bonus payments of $0.01, $0.05, or $0.10 per every 10 sliders they completed accurately. In order to complete a slider, workers needed to drag it with their mouse to a value indicated on the left side of the slider; sliders ranged from 1 to 300 (see Gill and Prowse 2011, for previous uses of the task to measure real effort). The indicated values were generated randomly between 1 and 300. Workers were given 5 minutes to complete up to 90 sliders ad libitum. Importantly, to receive payment, workers could only proceed to the next page of the study after 5 minutes, no matter how many sliders they decided to complete. To ensure that workers correctly understood their task, they could only start moving sliders after having answered a comprehension question correctly.

Forecasters were randomly assigned to one of four experimental treatments: a JP treatment and three SP treatments. All forecasters read a description of the slider task and practiced five slider tasks on their own. As a benchmark, we informed forecasters in all treatments that the highest worker performance had been 87 out of 90 sliders. As in Experiment 1A, forecasters in the JP treatment predicted the average number of sliders completed across the three wage levels while those in the SP treatments made predictions for one of the randomly assigned wage levels. Specifically, forecasters in the JP treatment were asked to predict "the average number of sliders correctly completed by participants in an online experiment who received varying levels of bonus payment in addition to the \$0.55 base rate compensation" whereas forecasters in the SP treatment were asked to predict "the average number of sliders correctly completed by participants in an online experiment who received a certain amount of bonus payment in addition to the \$0.55 base rate compensation." As in Experiment 1, forecasters received an \$0.50 bonus payment for accurate predictions in addition to their base payment of \$0.55.

Forecasters who acted as managers were given a description of the workers' slider task and were tasked with hiring workers from one of the three wage groups. Specifically, the managers were told, "as the employer, you can choose which wage group to hire the worker from. You will be paid according to the average per-worker performance of the group you hire, minus their per-slider wage," and were provided with the following profit function: P = (\$0.30 - Group's wage) × Average performance of the Group. Participants were then told that "the higher the wage, the less profit per slider you will make. However, higher payment may also be more motivating for the workers, so that they complete more sliders on average. Your task is to choose a wage that you think maximizes your profit." To facilitate profit calculations, we inserted an online calculator for managers to use. They then chose the group of workers from the three wage levels to hire from. Upon completion of the study, managers were paid their corresponding profit as a bonus payment in addition to their \$0.55 base rate payment.

Results: Following the pre-registered analysis plan, we excluded data from workers who failed to pass the comprehension check questions twice in a row (1 worker in the slider task, 178 forecasters, and 10 managers did so). Our final sample included the responses from 1,395 participants (300 workers, 987 forecasters, and 108 managers).

Workers' actual effort was not responsive to increases in wages (Difference in actual effort between \$0.01 and \$0.05 wages: $M = -0.510, t(196) = -0.251, p = 0.802, Cohen's\ d = -0.036$; Difference in actual effort between \$0.05 and \$0.10 wages: $M = -0.304, t(198) = -0.152, p = 0.880, Cohen's\ d = -0.022$).

How did forecasters' predictions compare to the workers' actual effort? Similar to Experiment 1A, forecasters in the JP treatment predicted that workers would strongly respond to increases in wages (Difference in predicted effort between \$0.01 and \$0.05 wages: $M = 8.856, t(470) = 4.787, p < 0.001, Cohen's\ d = 0.441$; Difference in predicted effort between \$0.05 and \$0.10 wages: $M = 10.424, t(470) = 5.819, p < 0.001, Cohen's\ d = 0.536$). Unlike forecasters in the JP treatment, forecasters in the SP treatment predicted a much more muted response to wage increases (Difference in predicted effort between \$0.01 vs. \$0.05 wages: $M = 1.800, t(494) = 1.079, p = 0.281, Cohen's\ d = 0.097$; Difference in predicted effort between \$0.05 vs. \$0.10 wages: $M = 2.843, t(498) = 1.795, p = 0.073, Cohen's\ d = 0.187$). Table 1, Column 2 shows that the forecasted wage response was significantly smaller in the SP treatment than the JP treatment.

Mirroring Experiment 1A, forecasters in the SP treatments were fairly accurate in predicting workers' sensitivity to wages, whereas forecasters in the JP treatments substantially overestimated workers' responsiveness. This difference in accuracy is illustrated in Figure 2, which depicts actual and predicted effort responses when increasing the wage from \$0.01 to \$0.05 (Panel A) and from \$0.05 to \$0.10 (Panel B). In both cases, forecasts in the SP treatments did not differ from actual effort responsiveness, whereas forecasts in the JP treatment predicted much greater sensitivity. Table A5 in Appendix A corroborates these differences in accuracy within a regression framework.

We now turn to the decisions of the managers. From the workers' actual effort, we can calculate the expected profit of hiring a worker from a specific wage group. It is straightforward to show that it is profit-maximizing to hire a worker from the lowest wage group of \$0.01. Beliefs in the SP treatments would predict this choice as well: if workers are predicted to be relatively insensitive to wage increases, then \$0.01 is the profit-maximizing wage choice. At the same time, if workers' performance responded to wage levels in line with beliefs in the JP treatment, then the middle wage of \$0.05 is profit maximizing in expectation.

Figure 2 (Panel C) depicts the wage choices of the managers. The majority (55%) indeed chose to hire workers from the \$0.05 wage level, consistent with beliefs elicited in the JP treatment. This led to significantly lower profits than those expected under the profit-maximizing wage level[5].

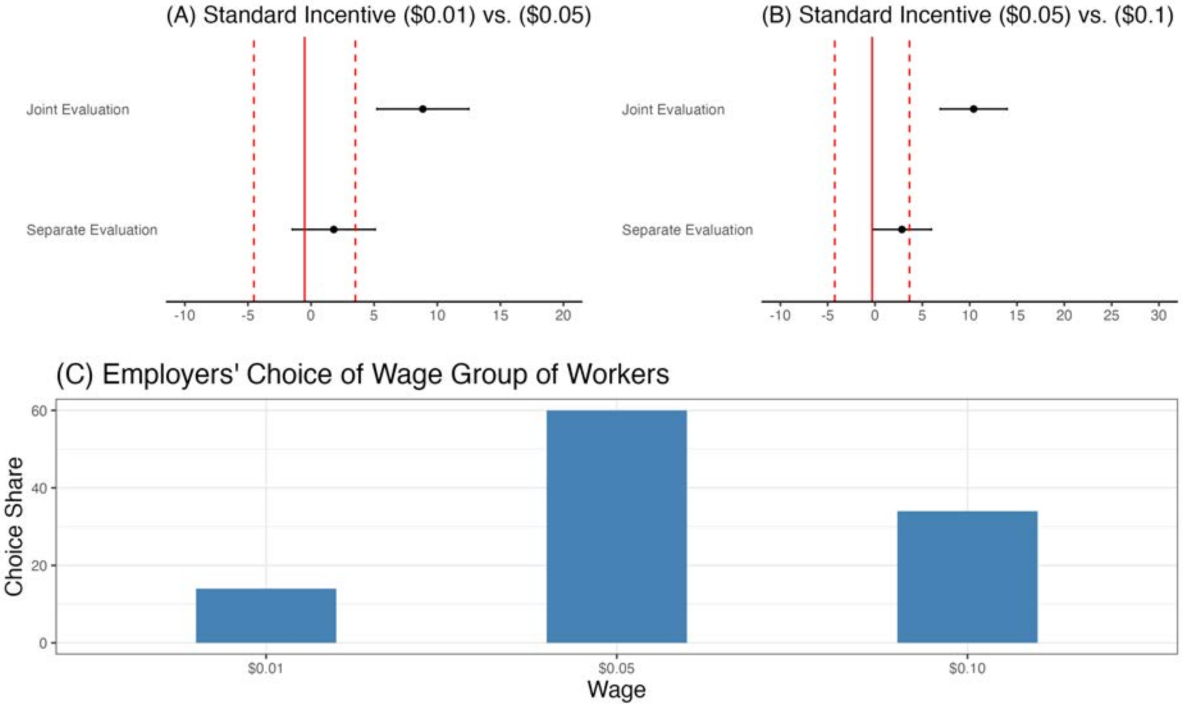


Figure 2. Forecasting Accuracy and Wage Choice in Experiment 1B

Notes: Panel A & B: The solid red lines indicate the observed differences in the average number of sliders completed by workers in the $0.01 and $0.05 (Panel A) and the $0.05 and $0.10 wage treatments (Panel B). The dashed lines indicate 95% confidence intervals, and the black dots indicate the predicted differences in the average number of sliders completed (whiskers indicate 95% confidence intervals). Panel C: Managers' choices of wage levels.

Our first two experiments showed that forecasters predicted greater responsiveness to treatment variation when in joint prediction mode than in separate prediction mode. The latter more accurately matched people's actual responses to the variation, both in a task where actual effort was sensitive to increases in wages (Experiment 1A) and in a task where actual effort was relatively insensitive (Experiment 1B). Moreover, forecast errors were consequential for decisions, leading managers in joint prediction mode to systematically deviate from the profit-maximizing benchmark.

[5] The expected profit at the profit-maximizing wage was $12.76. The average profit of managers in our experiment was $10.41, which is significantly less than the expected profit at the profit-maximizing wage (one-sample t-test, $t = -17.97, p < .001$).

3 Experiment 2

Our third experiment aims to provide evidence for the discrepancy of forecasts made in joint versus separate prediction mode in the context of predicting scientific findings. For this purpose, we build on the work of Dellavigna and Pope (2018), who conducted a large online study in which participants in the role of workers completed a real-effort task under different incentive levels, and expert forecasters predicted workers' performance under the different payment schemes. Building on this work, we collected data from participants in the role of workers and had them perform the task adopted from Dellavigna and Pope (2018) under four of the incentive levels tested in the original paper. We then asked participants recruited from the same platform to predict the workers' performance under the four incentive schemes, elicited in the JP or SP treatments.

Procedure: We recruited a total of 3,374 participants from Prolific and Mechanical Turk. Of these, 1,964 participants were recruited for the real-effort task and 1,225 were recruited for the forecasting task. Both samples were recruited and pre-registered separately.

Participants in the role of workers performed the same real-effort task as those in Dellavigna and Pope (2018), which consisted of successively entering the letter combination "ab." We instructed participants to alternately press the "a" and "b" buttons on their keyboard as quickly as possible for 10 minutes and explained that they would receive a point for each "ab" combination they entered. The instructions also explained that copy-pasting the letters into the text box was not possible.

Workers completed the task under either standard incentives that paid them a piece rate based on their performance (Standard Incentive treatments), or prosocial incentives that tied performance to charitable donations (Imas, 2014). Specifically, in the Standard $0.01 ($0.10) Incentive treatment, workers learned that "As a bonus, you will be paid an extra 1 (10) cents for every 100 points that you will score." In the Prosocial $0.01 ($0.10) Incentive treatment, workers' effort was tied to a charitable contribution, and workers learned that "The Red Cross will be given 1 (10) cent for every 100 points that you score."

Forecasters learned about the "ab" typing task and practiced it themselves before making predictions. Forecasters in the JP treatment were asked to predict the average amount of points earned by workers across all four incentive conditions; forecasters in SP treatments were randomly assigned to predict the average amount of points earned by workers in one of the four incentive conditions. The JP treatment thus follows the forecasting design of DellaVigna and Pope (2018), as well as other forecasting studies in the literature (see e.g., Dreber et al. 2015; Gordon et al. 2021)

We informed forecasters that "Individual scores for this task ranged from 0 to 5,246 points with only 8 people out of nearly 1,961 scoring above 4,000." In order to incentivize participants for accuracy, we followed DellaVigna and Pope (2018) and told participants that five participants would be randomly selected to win a bonus based on the accuracy of their forecasts, calculated as ($100 - mean squared error/2,000).

Results: Following the pre-registered analysis plan, we excluded data from workers who failed to pass the comprehension check questions (3 workers and 19 forecasters). Our final sample included the responses from 3,167 participants (1,961 workers and 1,206 forecasters).

Workers' effort choices did not significantly respond to increases in standard incentives (Difference in actual effort: $M = -48.0, t(986) = -1.25, p = 0.212$, $Cohen's\ d = -0.0794$) nor prosocial incentives (Difference in actual effort: $M = -70.4, t(971) = -1.75, p = 0.0812, Cohen's\ d = -0.112$).

Can forecasters predict workers' responses to both standard and psychological incentives? As in Experiments 1A and 1B, forecasters in the JP treatment predicted substantial effort responses to increases in standard incentives (Predicted difference: $M = 263.34, t(462) = 7.52, p < .001, Cohen's\ d = 0.698$) and prosocial incentives (Predicted difference: $M = 149.26, t(462) = 4.23, p < .001, Cohen's\ d = 0.393$)[6] The latter result is in line with Dellavigna and Pope (2018) who found that forecasters predicted significant effort differences in response to increases in prosocial incentives.

In contrast to forecasters in the JP treatment, those in the SP treatment predicted much more muted responses for both incentive types. For standard incentives, forecasters in SP treatment predicted a marginally significant increase in effort ($M = 53.18, t(491) = 1.67, p = 0.0961, Cohen's\ d = 0.15$); for prosocial incentives, they predicted a slight decrease in effort ($M = -81.07, t(479) = -2.60, p = 0.010, Cohen's\ d = -0.237$). As shown in Table 2, predicted effort responses in the SP treatments were significantly smaller across both incentive types compared to the JP treatment.

[6] The observed difference between the predicted effort under the low and high prosocial incentives in Dellavigna and Pope (2018) was remarkably similar ($M_{diff} = 103$) to the difference that we observed among our predictors ($M_{diff} = 149.26$).

Table 2. Forecasts in the JP and SP treatments in Experiment 2

	Prediction	
	(1)	(2)
	Standard Incentive	Prosocial Incentive
Constant	1 638.06***	1 489.03***
	(25.53)	(25.24)
$0.10 Incentive	263.34***	149.26***
	(35.00)	(35.30)
Separate Prediction	105.06***	299.38***
	(34.57)	(32.97)
$0.10 Incentive × Separate Prediction	-210.17***	-230.33***
	(47.36)	(47.08)
Observations	957	945

* p < 0.1, ** p < 0.05, *** p < 0.01

Notes: This table reports results from OLS regression, with robust standard errors in parentheses. The outcome measure is predicted performance scores. $0.10 Incentive is a binary indicator coded as 1 when the incentive was $0.10 and 0 when it was $0.01. Joint Prediction Mode is a is a binary indicator coded for forecasting data in the JP treatment and 0 for forecasting data in the SP treatment. Data from the JP treatments is treated as statistically independent to make it comparable to the forecasts in the SP treatments to observed (actual) effort.

Figure 3 shows that forecasters in the SP treatments were relatively accurate in predicting workers' sensitivity to standard (Panel A) and prosocial incentives (Panel B); neither estimate differs significantly from actual effort responses. In contrast, forecasters in the JP treatment substantially overestimated workers' responsiveness to both standard and prosocial incentives. Table B2 in Appendix B presents regression results corroborating this difference in forecasting accuracy. Together, the findings from Experiment 2 suggest that prediction mode is an important determinant of accuracy in the forecasting of scientific results.

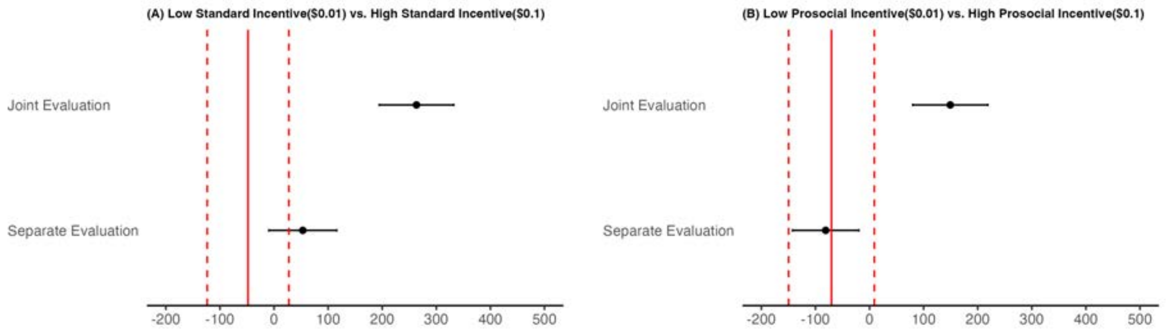


Figure 3. Forecasting Accuracy in Experiment 2

Notes: The solid red lines indicate the observed differences in the average number of points achieved by workers in the $0.01 and $0.10 standard (Panel A) and the $0.01 and $0.10 prosocial incentive treatments (Panel B). The dashed lines indicate 95% confidence intervals, and the black dots indicate the predicted differences in the average number of points achieved (whiskers indicate 95% confidence intervals).

4 Discussion

This paper examines how joint versus separate prediction mode impacts forecasts of causal effects. Three experiments show that when people evaluate the potential impact of variables jointly, they predict larger differences in outcomes than when the same predictions are made for each variable in isolation. In the first two experiments, forecasters in joint prediction mode predicted much larger worker responses to wage increases than those in separate prediction mode, and the latter were much closer to workers' actual responses. We show that this difference in forecasting accuracy can be costly, as people choosing wages in joint prediction mode overshot the profit-maximizing option. We also explore the implications of separate versus joint predictions for the forecasting of scientific results. Forecasters in joint prediction mode predicted larger treatment effects than those in the separate prediction mode, and were less accurate in their predictions of responses to both standard and psychological incentives.

While most of our evidence suggests that separate prediction mode—which more closely aligns with the decision environment of agents—leads to more well-calibrated forecasts of behavioral responses, this may not always be the case. The accuracy consequences of prediction modes likely depend on the mental model of the forecaster: separate prediction mode is likely to generate more accurate forecasts when agents make decisions in response to one variable realization. When agents are experienced or are given information about multiple variable realizations, joint prediction mode is likely to produce more accurate forecasts. As such, while joint predictions will almost always lead to larger differences in outcome forecasts than separate predictions, the accuracy consequences depend on the agents' experience, the extent to which they engage in counterfactual thinking, and the decision environment.

Additionally, there may be other benefits for using joint prediction mode that lead it to be used for forecasting in practice. One important factor is power. Joint prediction mode allows one to elicit forecasts across all realizations of interest from one individual; separate prediction mode naturally requires recruiting different individuals to make forecasts for each realization. In a setting where, for example, a researcher is interested in eliciting expert forecasts and the expert pool is small, the use of separate prediction mode may not be feasible. Additionally, it may be the case that the researcher is only interested in forecasting the ranking of interventions rather than the size of behavioral responses. Joint forecasting mode may be more accurate for the former than the latter.

Future research should explore the psychological mechanism that generates differences between prediction modes. While our results suggest that joint versus separate prediction modes lead to differences in mental representations of counterfactual variable responses, the specific channel that generates forecast differences is unclear. One potential mechanism is a contrast effect (Pepitone and DiNubile, 1976; Hartzmark and Shue, 2018), where forecasters project salient differences in variable levels to differences in the resulting outcomes. Another possibility may be that joint versus separate prediction mode prompts forecasters to use different scales when predicting responses (Goldstein and Einhorn, 1987). For example, while people in the joint prediction mode may view the given differences in variables as meaningful, those in the separate prediction mode require larger differences to generate the same response. Finally, forecasters in joint prediction mode may simulate a different set of agents' preferences than those in the separate prediction mode—a phenomenon termed distinction bias (Hsee and Zhang, 2004). A better understanding of the underlying mechanism has the potential to illuminate which prediction mode will lead to more accurate forecasts in a given context.

References

Ariely, D., G. Loewenstein, and D. Prelec (2003): “"Coherent Arbitrariness": Stable Demand Curves without Stable Preferences,” *The Quarterly Journal of Economics*, 118, 73–105.

Bazerman, M. H., G. F. Loewenstein, and S. B. White (1992): “Reversals of Preference in Allocation Decisions: Judging an Alternative Versus Choosing Among Alternatives,” *Administrative Science Quarterly*, 37, 220–240.

Bessone, P., F. R. Campante, C. Ferraz, and P. Souza (2022): “Social Media and the Behavior of Politicians: Evidence from Facebook in Brazil,” Report 30306, National Bureau of Economic Research.

Blount, S. and M. H. Bazerman (1996): “The Inconsistent Evaluation of Absolute Versus Comparative Payoffs in Labor Supply and Bargaining,” *Journal of Economic Behavior & Organization*, 30, 227–240.

Bohnet, I., A. van Geen, and M. Bazerman (2016): “When Performance Trumps Gender Bias: Joint vs. Separate Evaluation,” *Management Science*, 62, 1225–1234.

Bohren, J. A. and D. N. Hauser (2021): “Learning With Heterogeneous Misspecified Models: Characterization and Robustness,” *Econometrica*, 89, 3025–3077.

Bordalo, P., N. Gennaioli, Y. Ma, and A. Shleifer (2020): “Overreaction in Macroeconomic Expectations,” *American Economic Review*, 110, 2748–2782.

Bowen, D. (2022): “Simple Models Predict Behavior at Least as Well as Behavioral Scientists,” .

Charness, G., U. Gneezy, and M. A. Kuhn (2012): “Experimental methods: Between-subject and within-subject design,” *Journal of economic behavior & organization*, 81, 1–8.

DellaVigna, S. and E. Linos (2022): “RCTs to Scale: Comprehensive Evidence From Two Nudge Units,” *Econometrica*, 90, 81–116.

DellaVigna, S. and D. Pope (2018): “Predicting Experimental Results: Who Knows What?” *Journal of Political Economy*, 126, 2410–2456.

Dellavigna, S. and D. Pope (2018): “What Motivates Effort? Evidence and Expert Forecasts,” *Review of Economic Studies*, 85, 1029–1069.

DellaVigna, S., D. Pope, and E. Vivalt (2019): “Predict Science to Improve Science,” *Science*, 366, 428–429.

DREBER, A., T. PFEIFFER, J. ALMENBERG, S. ISAKSSON, B. WILSON, Y. CHEN, B. A. NOSEK, AND M. JOHANNESSON (2015): "Using prediction markets to estimate the reproducibility of scientific research," *Proceedings of the National Academy of Sciences of The United States of America*, 112, 15343–15347.

FOX, C. R. AND A. TVERSKY (1995): "Ambiguity aversion and comparative ignorance," *The quarterly journal of economics*, 110, 585–603.

FREDERICK, S. AND B. FISCHHOFF (1998): "Scope (in) sensitivity in elicited valuations," *Risk Decision and Policy*, 3, 109–123.

GENNAIOLI, N. AND A. SHLEIFER (2010): "What Comes to Mind," *Quarterly Journal of Economics*, 125, 1399–1433.

GILL, D. AND V. L. PROWSE (2011): "A Novel Computerized Real Effort Task Based on Sliders," *Available at SSRN 1732324*.

GOLDSTEIN, W. M. AND H. J. EINHORN (1987): "Expression Theory and The Preference Reversal Phenomena," *Psychological Review*, 94, 236–254.

GORDON, M., D. VIGANOLA, A. DREBER, M. JOHANNESSON, AND T. PFEIFFER (2021): "Predicting Replicability-Analysis of Survey and Prediction Market Data from Large-Scale Forecasting Projects," *Plos One*, 16.

HARTZMARK, S. M. AND K. SHUE (2018): "A Tough Act to Follow: Contrast Effects in Financial Markets," *Journal of Finance*, 73, 1567–1613.

HSEE, C. K. (1996): "The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives," *Organizational Behavior and Human Decision Processes*, 67, 247–257.

HSEE, C. K., G. F. LOEWENSTEIN, S. BLOUNT, AND M. H. BAZERMAN (1999): "Preference Reversals Between Joint and Separate Evaluations of Options: A Review and Theoretical Analysis," *Psychological Bulletin*, 125, 576–590.

HSEE, C. K. AND J. ZHANG (2004): "Distinction bias: misprediction and mischoice due to joint evaluation." *Journal of personality and social psychology*, 86, 680.

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

MEDVEC, V. H., S. F. MADEY, AND T. GILOVICH (1995): "When Less Is More: Counterfactual Thinking and Satisfaction Among Olympic Medalists," *Journal of Personality and Social Psychology*, 69, 603–610.

Milkman, K. L., D. Gromet, H. Ho, J. S. Kay, T. W. Lee, P. Pandiloski, Y. Park, et al. (2021a): "Megastudies Improve the Impact of Applied Behavioural Science," *Nature*, 600, 478–483.

Milkman, K. L., M. S. Patel, L. Gandhi, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, et al. (2021b): "A Megastudy of Text-Based Nudges Encouraging Patients to Get Vaccinated at an Upcoming Doctor's Appointment," *Proceedings of the National Academy of Sciences*, 118, e2101165118.

Mishkin, M., L. G. Ungerleider, and K. A. Macko (1983): "Object Vision and Spatial Vision: Two Cortical Pathways," *Trends in Neurosciences*, 6, 414–417.

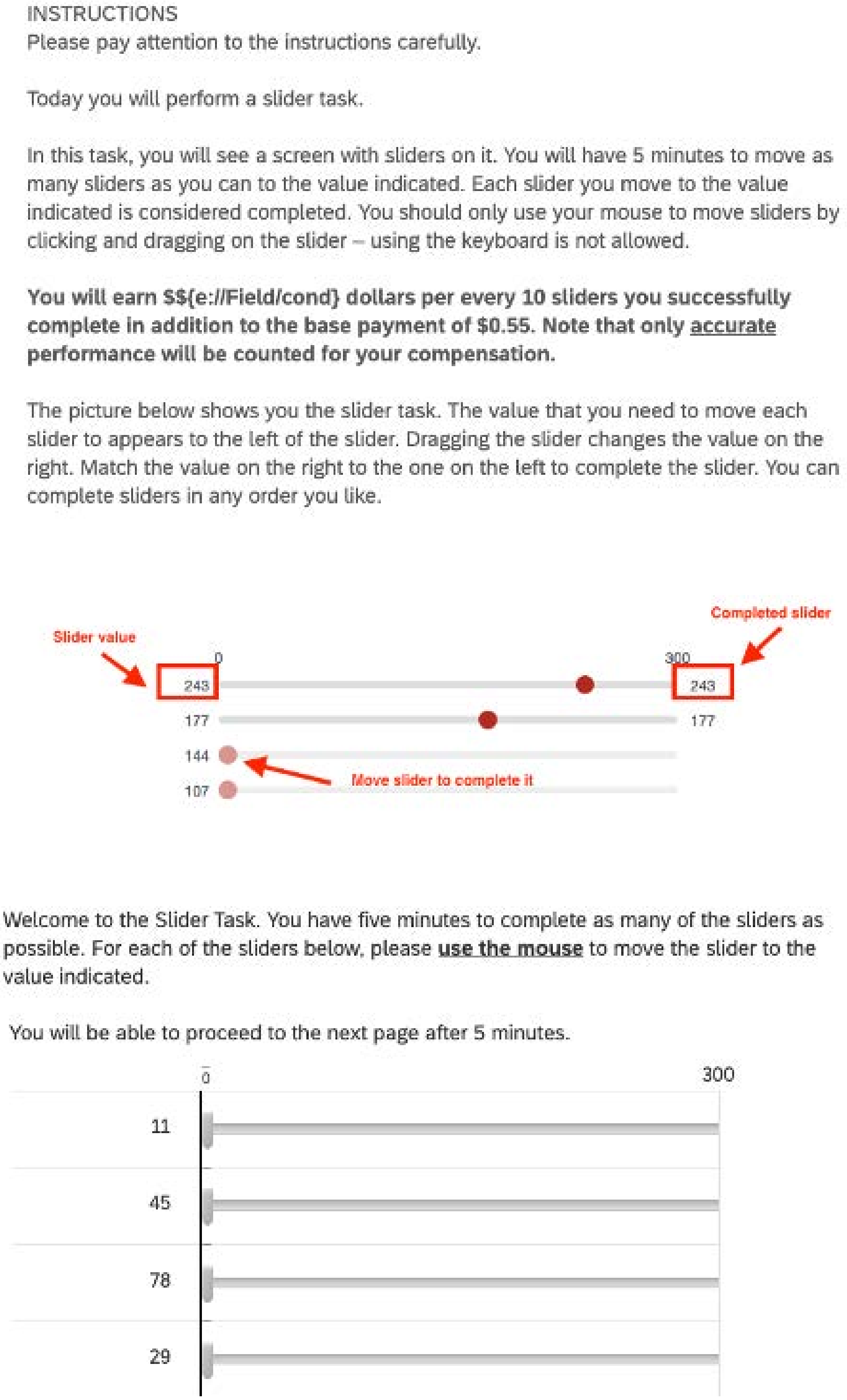
Pepitone, A. and M. DiNubile (1976): "Contrast Effects in Judgments of Crime Severity and the Punishment of Criminal Violators," *Journal of Personality and Social Psychology*, 33, 448–459.

Roese, N. J. (1997): "Counterfactual Thinking," *Psychological Bulletin*, 121, 133–148.

Schwartz, D., E. A. Keenan, A. Imas, and A. Gneezy (2021): "Opting-In to Prosocial Incentives," *Organizational Behavior and Human Decision Processes*, 163, 132–141.

Appendix A. APPENDIX A: INSTRUCTIONS AND ADDITIONAL ANALYSIS FOR EXPERIMENTS 1A AND 1B

A.1 INSTRUCTIONS



INSTRUCTIONS
Please pay attention to the instructions carefully.

Today you will perform a slider task.

In this task, you will see a screen with sliders on it. You will have 5 minutes to move as many sliders as you can to the value indicated. Each slider you move to the value indicated is considered completed. You should only use your mouse to move sliders by clicking and dragging on the slider – using the keyboard is not allowed.

You will earn $${e://Field/cond} dollars per every 10 sliders you successfully complete in addition to the base payment of $0.55. Note that only accurate performance will be counted for your compensation.

The picture below shows you the slider task. The value that you need to move each slider to appears to the left of the slider. Dragging the slider changes the value on the right. Match the value on the right to the one on the left to complete the slider. You can complete sliders in any order you like.

Welcome to the Slider Task. You have five minutes to complete as many of the sliders as possible. For each of the sliders below, please use the mouse to move the slider to the value indicated.

You will be able to proceed to the next page after 5 minutes.

Figure A.1. Experiment 1a: LINK SEARCH TASK INSTRUCTIONS

A.2 ADDITIONAL ANALYSES FOR EXPERIMENTS 1A AND 1B

Table A.1. DESCRIPTIVE STATISTICS OF ACTUAL AND PREDICTED EFFORT ON THE IMAGE COLLECTION TASK IN TWO WAGE GROUPS in EXPERIMENT 1A

Type	Wage	N	Mean	SD	SE	Median
Actual	\$0.02	435	13.25	12.38	0.59	10
	\$0.10	476	15.82	12.30	0.56	15
Separate Prediction	\$0.02	399	18.90	7.95	0.40	21
	\$0.10	397	20.76	6.99	0.35	21
Joint Prediction	\$0.02	398	14.21	7.78	0.39	13
	\$0.10	398	22.65	6.67	0.33	24

Table A.2. TREATMENT DIFFERENCES IN ACTUAL AND PREDICTED EFFORT LEVELS IN EXPERIMENT 1A

Treatment Difference \$0.02 - \$0.10 Incentive		
Actual Behavior	Separate Prediction	Joint Prediction
$M_{\text{diff}}(CI_{\text{diff}}), N$		
2.387(0.822, 3.951), 911	1.853(0.812, 2.895), 796	8.442(7.434, 9.451), 796
t=2.994, p=0.003	t=3.494, p<0.001	t=16.434, p<0.001

Notes: t-tests are independent tests. In all our analyses, we treat data from the joint predictions as statistically independent so they are comparable to the separate predictions and observed (actual) effort levels.

Table A.3. DESCRIPTIVE STATISTICS OF ACTUAL AND PREDICTED EFFORT ON THE SLIDER TASK IN THREE WAGE GROUPS in EXPERIMENT 1B

	Wage	N	Mean Effort (SD)	Median Effort	Mean Profit (SD)
Actual Effort (N = 300)	\$0.01	100	44.00 (13.81)	45.5	12.76 (4.00)
	\$0.05	98	43.49 (14.83)	42	10.87 (3.71)
	\$0.10	102	43.19 (13.46)	43	8.64 (2.69)
	Wage	N	Mean Effort (SD)	Median Effort	Calculated Average Profit (SD)
Predicted Effort in Separate Predictions (N = 751)	\$0.01	251	55.50 (18.85)	56	16.10 (5.47)
	\$0.05	245	57.30 (18.31)	60	14.33 (4.58)
	\$0.10	255	60.15 (17.10)	60	12.03 (3.42)
	Wage	N	Mean Effort (SD)	Median Effort	Calculated Average Profit (SD)
Predicted Effort in Joint Predictions (N = 236)	\$0.01	236	44.93 (21.29)	43	13.03 (6.17)
	\$0.05	236	53.79 (18.82)	55	13.45 (4.71)
	\$0.10	236	64.21 (20.08)	68.5	12.84 (4.02)
	Wage	N	Choice Share (%)		Mean Profit (SD)
Wage Choice (N = 108)	\$0.01	14	12.96	-	-
	\$0.05	60	55.56	-	-
	\$0.10	34	31.48	-	-

Table A.4. TREATMENT DIFFERENCES IN ACTUAL AND PREDICTED EFFORT LEVELS IN EXPERIMENT 1B

Treatment Difference $0.01 - $0.05 Incentive		
Actual Behavior	Separate Prediction	Joint Prediction
$M_{\text{diff}}(CI_{\text{diff}}), N$		
-0.510(-4.526, 3.505), 198	1.800(-1.479, 5.079), 496	8.856(5.221, 12.491), 472
t=-0.251, p=0.802	t=1.079, p=0.281	t=4.787, p<0.001
Treatment Difference $0.05 - $0.10 Incentive		
Actual Behavior	Separate Prediction	Joint Prediction
$M_{\text{diff}}(CI_{\text{diff}}), N$		
-0.304(-4.251, 3.643), 200	2.843(-0.268, 5.954), 500	10.424(6.903, 13.944), 472
t=-0.152, p=0.880	t=1.795, p=0.073	t=5.819, p<0.001

Notes: t-tests are independent tests. In all our analyses, we treat data from the joint predictions as statistically independent so they are comparable to the separate predictions and observed (actual) effort levels.

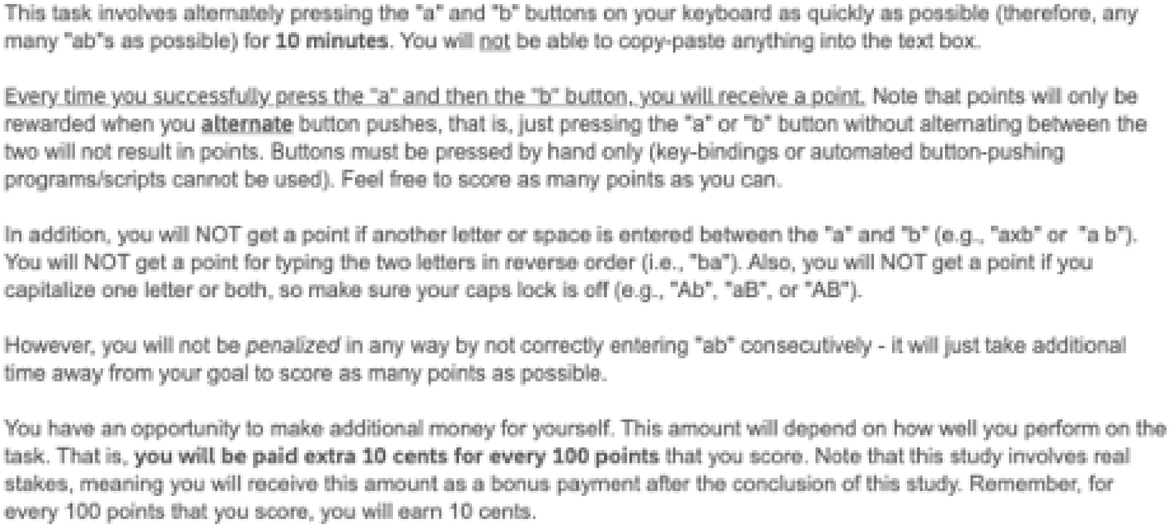
Table A.5. PREDICTION ACCURACY: EXPERIMENT 1A AND EXPERIMENT 1B

	Accuracy	
	Experiment 1A	Experiment 1B
	(1)	(2)
Constant	13.25***	44.00***
	(0.59)	(1.38)
Joint Prediction	0.96	0.93
	(0.71)	(1.95)
Separate Prediction	5.65***	11.50***
	(0.71)	(1.82)
$0.1 Incentive	2.57***	-0.81
	(0.82)	(1.91)
$0.1 Incentive × Joint Prediction	5.87***	20.09***
	(0.97)	(2.70)
$0.1 Incentive × Separate Prediction	-0.72	5.46**
	(0.98)	(2.50)
$0.05 Incentive		-0.51
		(2.03)
$0.05 Incentive × Joint Prediction		9.37***
		(2.75)
$0.05 Incentive × Separate Prediction		2.31
		(2.63)
Observations	2503	1759
$0.1 Incentive × Joint = $0.05 Incentive × Joint		p<0.001
$0.1 Incentive × Separate = $0.05 Incentive × Separate		p=0.218

Notes:This table reports results from OLS regression, focusing on actual effort and the corresponding forecasts in experiments 1A and 1B. The outcome measure is actual/predicted effort levels. $0.10 Incentive is a binary indicator coded as 1 when the incentive was $0.10 and 0 otherwise. $0.05 Incentive is a binary indicator coded as 1 when the incentive was $0.05 and 0 otherwise. Joint Prediction Mode is a binary indicator coded as 1 for JP and 0 otherwise. Separate Prediction Mode is a is a binary indicator coded as 1 for SP and 0 otherwise. The accuracy of forecasts is hence indicated by the product term of the wage and prediction mode dummies. Robust standard errors are reported in parenthesis.

Appendix B. APPENDIX B: INSTRUCTIONS AND ADDITIONAL ANALYSES FOR EXPERIMENT 2

B.1 INSTRUCTIONS (Low Standard incentive Treatment)



This task involves alternately pressing the "a" and "b" buttons on your keyboard as quickly as possible (therefore, any many "ab"s as possible) for **10 minutes**. You will not be able to copy-paste anything into the text box.

Every time you successfully press the "a" and then the "b" button, you will receive a point. Note that points will only be rewarded when you **alternate** button pushes, that is, just pressing the "a" or "b" button without alternating between the two will not result in points. Buttons must be pressed by hand only (key-bindings or automated button-pushing programs/scripts cannot be used). Feel free to score as many points as you can.

In addition, you will NOT get a point if another letter or space is entered between the "a" and "b" (e.g., "axb" or "a b"). You will NOT get a point for typing the two letters in reverse order (i.e., "ba"). Also, you will NOT get a point if you capitalize one letter or both, so make sure your caps lock is off (e.g., "Ab", "aB", or "AB").

However, you will not be *penalized* in any way by not correctly entering "ab" consecutively - it will just take additional time away from your goal to score as many points as possible.

You have an opportunity to make additional money for yourself. This amount will depend on how well you perform on the task. That is, **you will be paid extra 10 cents for every 100 points** that you score. Note that this study involves real stakes, meaning you will receive this amount as a bonus payment after the conclusion of this study. Remember, for every 100 points that you score, you will earn 10 cents.

Figure B.1. Experiment 2: INSTRUCTIONS

Table B.1. DIFFERENCES BETWEEN ACTUAL AND PREDICTED EFFORT LEVELS IN STANDARD AND PROSOCIAL INCENTIVES TREATMENTS

Standard Incentive		
Difference in Actual/Predicted Effort Levels between High and Low Standard Incentive		
Actual Behavior	Separate Prediction	Joint Prediction
$M_{\text{diff}}(CI_{\text{diff}}), N$		
-48.042(-123.557, 27.473), 988 t=-1.248, p=0.212	53.175(-9.498, 115.849), 493 t=1.667, p=0.096	263.341(194.551, 332.130), 464 t=7.523, p<0.001
Prosocial Incentive		
Difference in Actual/Predicted Effort Levels between High and Low Prosocial Incentive		
Actual Behavior	Separate Prediction	Joint Prediction
$M_{\text{diff}}(CI_{\text{diff}}), N$		
-70.418(-149.569, 8.734), 973 t=-1.746, p=0.081	-81.065(-142.273, -19.857), 481 t=-2.602, p=0.010	149.263(79.892, 218.633), 464 t=4.228, p<0.001

Notes: (i) In the standard incentive treatments, M_{diff} (CI_{diff}) indicates the mean and standard deviation of the mean effort levels between the high and low standard incentive treatments. In the prosocial treatments, it means the mean and standard deviation of the mean effort levels between the high and low prosocial incentive treatments. (ii) t-tests were independent t-tests. In all our analyses, we treat data from the joint predictions as statistically independent so they are comparable to the separate predictions and observed (actual) effort levels. Forecasters in the Joint Prediction Mode treatment predicted that increasing incentives

Table B.2. PREDICTION ACCURACY

	Accuracy	
	(1)	(2)
	Standard Incentive	Prosocial Incentive
Constant	1 656.64***	1 511.12***
	(29.15)	(27.04)
$0.10 Incentive	-48.04	-70.42*
	(38.57)	(40.40)
Joint Prediction Mode	-18.59	-22.09
	(38.74)	(36.98)
Separate Prediction Mode	86.47**	277.30***
	(37.32)	(34.36)
$0.10 Incentive × Joint Prediction Mode	311.38***	219.68***
	(52.07)	(53.64)
$0.10 Incentive × Separate Prediction Mode	101.22**	-10.65
	(50.04)	(51.01)
Observations	1945	1918
Incentive × Joint = Incentive × Separate	P<0.001	P<0.001

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports results from OLS regression, focusing on actual effort data from the "ab" typing study and the corresponding forecasts from the Forecasting study. The outcome measure is actual/predicted effort levels. $0.10 Incentive is a binary indicator coded as 1 when the incentive was $0.10 and 0 otherwise. Joint Prediction Mode is a binary indicator coded as 1 for JP and 0 otherwise. Separate Prediction Mode is a is a binary indicator coded as 1 for SP and 0 otherwise. The accuracy of forecasts is hence indicated by the product term of the wage and prediction mode dummies. Robust standard errors are reported in parenthesis.