Pre-analysis plan:
The Dark Side of Overconfidence

Christoph Drobner* A. Yeşim Orhun†

December 9, 2023

1 Introduction

We study how (miscalibrated) beliefs about one's ability causally affect optimal actions through (misguided) inferences about an external fundamental. Our experimental paradigm ensures that the causal impact of beliefs on effort solely operates through inferences about an external fundamental to eliminate any motivational confounds arising from learning about one's ability.

2 Experimental Design

Our experimental design consists of two periods. Participants will receive a completion fee of $2 and the bonus payment from one randomly chosen period of the experiment. We plan to collect data from 2,000 US participants on *Prolific*.

2.1 Period 1

In period 1, participants perform a logic quiz with 12 puzzles from Civelli et al. (2018) that are similar to the Raven Progressive Matrix test (a commonly used test to measure fluid intelligence). In each puzzle, participants have 20 seconds to choose the correct answer from a set of four possible answers. Participants earn an additional $0-$4 bonus if period 1 is chosen to determine payments, and know that the chance of earning higher amounts depends on their performance in the logic quiz.

*Technical University Munich, UniDistance Suisse; Email: christoph.drobner@tum.de
†University of Michigan; Email: aorhun@umich.edu

In a between-subjects design, we randomly assign participants to easy and difficult versions of the logic quiz. This is the first of two randomized treatments in the design. We refer to the two conditions as Easy Quiz and Difficult Quiz from here on. While questions 1-7 are the same among both conditions, questions 8-12 vary in difficulty levels between the Easy Quiz and Difficult Quiz conditions. The purpose of this manipulation is to provide exogenous variation in participants' beliefs about their relative performance in the logic quiz. Based on Moore and Healy (2008), we expect participants to form more optimistic beliefs about their relative performance in the easy version of the logic quiz compared to the difficult version of the logic quiz.

After finishing the logic quiz, participants learn their logic quiz score and indicate the likelihood of having scored in the top half in a group of 4 participants, including themselves. We refer to these beliefs as *performance beliefs* from here on. We incentivize the performance belief reports with the binarized scoring rule without providing detailed information about the incentives (as in Danz et al. (2022)). Participants know that they have the chance to win a $1 bonus, and the chance of winning this bonus increases with the accuracy of their performance belief reports.

The study then details period 1 payments. Participants learn that they receive two payoffs. One payoff is provided by evaluator 1, and the other by evaluator 2. Here, the second randomized treatment of the experimental design takes effect. We randomly assign (between-subjects) either the performance evaluator as evaluator 1 (and random evaluator as evaluator 2) or random evaluator as evaluator 1 (and performance evaluator as evaluator 2). The performance evaluator pays participants based on their relative performance in the logic quiz: if the participant's performance ranks in the top half of the group of 4 participants, the performance evaluator pays $2, and otherwise pays $0. In contrast, the random evaluator determines participants' payoff by tossing a coin: if the coin toss results in heads, the random evaluator pays them $2, and otherwise pays $0. Participants know that exactly one of the two evaluators is a performance evaluator and exactly one of the two evaluators is a random evaluator. However, participants are not aware of which evaluator corresponds to each role.

Consequently, participants receive one of four types of payoff information based on their logic quiz performance and the randomly assigned evaluator: Both High, Mixed 1, Mixed 2, Both Low. Participants who receive only high payoffs or low payoffs from both evaluators learn whether their performance in the logic quiz ranks among the top half and therefore are redirected to a different survey. Participants who receive a mixed set of payoffs from the two evaluators (i.e., one paying $2 and the other paying $0) do not learn about their rank in the logic quiz from the payoffs. These participants proceed to period 2.

2.2 Period 2

In period 2, participants are told that they are about to be asked to solve up to 25 decoding tasks and are presented with an example of the decoding task. The decoding task presents a panel with numbers and letters. Participants' task is to decode text from a 5 digit number and enter the answer in an input field. Participants are informed that they can decide whether they want to continue or stop working after each decoding task. This stopping option is introduced to increase the responsiveness of effort provision to monetary incentives by making the opportunity cost of working on the decoding task salient.

Participants are informed that they are paid for their work in period 2 by the same evaluator 1 from period 1. If evaluator 1 is the random evaluator, they receive no payoff for their work in period 2. If evaluator 1 is the performance evaluator, they receive \$0.1 for each correctly solved decoding task. We deliberately choose a piece-rate instead of a tournament-based performance payoff in period 2. This design feature ensures that participants' effort provision in period 2 depends on their beliefs about the returns to effort, but eliminates confounds arising from strategic considerations regarding other's beliefs about the returns to effort and their effort provisions. Before participants start working on the decoding tasks, we elicit their beliefs about the likelihood that evaluator 1 is the performance evaluator. Henceforth, we refer to these beliefs as *returns to effort beliefs* regarding the period 2 task. After the returns to effort belief elicitation, participants work on the 25 decoding tasks.

2.3 Randomization

The randomization of quiz difficulty is stratified across gender. The randomization of the evaluator assignment is stratified across quiz difficulty conditions, deciles of performance beliefs, gender, outcome of the coin toss and whether or not the participant's performance is in the top half of the score distribution.

2.4 Pre-Test

We conducted a pre-test of performance beliefs to calibrate the Easy Quiz and Difficult Quiz conditions such that participants form on average overconfident performance beliefs, and more overconfident beliefs in the Easy Quiz condition compared to the Difficult Quiz condition.

3 Theoretical Predictions

When individuals cannot separately identify the contribution of their ability and other relevant fundamentals to their productivity determined by the environment, Heidhues et al. (2018) and Hestermann and Le Yaouanq (2021) theoretically show that individuals may learn misguidedly about the environment, when they hold biased beliefs about their ability. This misguided learning is consistent with Bayesian updating and particularly strong for overconfident individuals when their beliefs about their ability do not converge to the truth (Heidhues et al., 2018).

Our experiment parallels this context with three design features. Recall that due to the random assignment of evaluator type, participants in the Mixed 1 group receive higher period 1 payoffs from evaluator 1, and participants in the Mixed 2 group receive higher period 1 payoffs from evaluator 2. The first design feature is that these payoff signals are jointly determined by ability (performance in period 1 logic quiz) and environment (by whether or not evaluator 1 is a performance evaluator). The second design feature is that, as we show below, these payoff signals do not contain information about ability. Therefore, we rule out potential confounding effects that learning ego-relevant information may have on future performance. Finally, we introduce exogenous variation in the degree of overconfidence by using easy and difficult versions of the logic quiz (Moore and Healy, 2008).

Given these three design features, models proposed by Heidhues et al. (2018) and Hestermann and Le Yaouanq (2021) would predict overconfident participants in the Mixed 1 group to (misguidedly) infer higher returns to effort in period 2 compared to participants in the Mixed 2 group. The difference between Mixed 1 and Mixed 2 groups would be increasing in the degree of overconfidence, which we expect to be higher in the Easy Quiz compared to the Difficult Quiz condition. In what follows, we formalize these claims.

No learning about ability from payment levels Participants in our experiment form beliefs $\gamma \in (0, 1)$ about the likelihood of having performed in the top half (H) in the logic quiz (i.e., $\gamma \stackrel{\text{def}}{=\!=} Pr(H)$). For exposition, denote γ_0 as their prior performance beliefs before observing payoff signals. Participants observe a payoff signal about scoring in the top half: one from evaluator 1 and one from evaluator 2. All participants in the Mixed 1 group received $(s_1 = H, s_2 = L)$; and all participants in the Mixed 2 group received $(s_1 = L, s_2 = H)$. Recall that the probability of evaluator 1 being the random evaluator is 0.5. By definition, $Pr(s_i = H|H) = 1$ when the evaluator i is a performance evaluator, and $Pr(s_i = H|H) = .5$ when the evaluator i is a random evaluator. Importantly, participants do not know which evaluator is which type. Thus, for any evaluator i, $Pr(s_i = H|H) = 0.75$

and $Pr(s_i = L|H) = 0.25$, resulting in $Pr(s_i = H, s_{-i} = L|H) = 0.5$. Consequently,

$$\begin{aligned}\gamma_1(s_i = H, s_{-i} = L) &= \frac{Pr(s_i = H, s_{-i} = L|H)\gamma_0}{Pr(s_i = H, s_{-i} = L|H)\gamma_0 + (1 - Pr(s_i = H, s_{-i} = L|H))(1 - \gamma_0)} \\ &= \frac{.5\gamma_0}{.5\gamma_0 + .5(1 - \gamma_0)} = \gamma_0\end{aligned} \tag{1}$$

Thus, after seeing a mixed payoff signal, the Bayesian posterior performance belief γ_1 about having performed in the top half (H) does not differ from the prior γ_0, i.e. $\gamma_0 = \gamma_1$, which we simply refer to as γ from here on. This feature of our experimental design rules out confounds that may arise due to learning about ability from mixed payoff signals. As we detail next, given beliefs γ on their relative performance, participants in MIXED 1 and MIXED 2 groups learn instead about the likelihood that evaluator 1 is the performance evaluator. Importantly, this learning process depends on whether the high signal (payoff) is coming from evaluator 1 (MIXED 1 group) or from evaluator 2 (MIXED 2 group).

Misguided learning about returns to effort Participants in our experiment form beliefs $\theta \in (0, 1)$ about the likelihood that evaluator 1 is the performance evaluator (P) (i.e., $\theta \stackrel{\text{def}}{=\!=} Pr(P)$). A priori, there is a 50% chance that evaluator 1 is the performance evaluator (P), i.e., $\theta_0 = .5$. Participants observe one signal (payoff) from evaluator 1, $s_1 \in \{H, L\}$. The information extracted from this signal about the likelihood that the evaluator 1 is the performance evaluator depends on participants' performance beliefs $\gamma \in (0, 1)$: $Pr(s_1 = H|P) = \gamma$ and $Pr(s_1 = L|P) = 1 - \gamma$. Therefore, the Bayesian posterior beliefs about the likelihood that evaluator 1 is the performance evaluator (P) when $s_1 = H$ (i.e., MIXED 1 group) and when $s_1 = L$ (i.e., MIXED 2 group) are equal to γ and $1 - \gamma$, respectively:

$$\begin{aligned}\theta_1(s_1 = H) &= \frac{pr(P)Pr(s_1 = H|P)}{pr(P)Pr(s_1 = H|P) + (1 - pr(P))(1 - Pr(s_1 = H|P))} \\ &= \frac{.5\gamma}{.5\gamma + .5(1 - \gamma)} = \gamma\end{aligned} \tag{2}$$

$$\begin{aligned}\theta_1(s_1 = L) &= \frac{pr(P)Pr(s_1 = L|P)}{pr(P)Pr(s_1 = L|P) + (1 - pr(P))(1 - Pr(s_1 = L|P))} \\ &= \frac{.5(1 - \gamma)}{.5(1 - \gamma) + .5\gamma} = 1 - \gamma\end{aligned} \tag{3}$$

As such, in line with Heidhues et al. (2018) and Hestermann and Le Yaouanq (2021), our experimental setup allows performance beliefs γ to impact inferences about the returns to effort from signals of productivity that are jointly determined by one's ability and the environment. Importantly, this inference is misguided to the extent that performance beliefs γ depart from one's actual ability. We calculate the objective probability that a participant scores in the top half of a randomly drawn comparison group of four individuals based on the participant's logic quiz score. We define *prior bias* as the difference between performance beliefs and the objective probability of ranking in the top half. We define *misguidedness* as the difference between the participant's returns to effort belief and the objective returns to effort belief (which is calculated by applying Bayes' rule to the objective probability of that individual ranking in the top half). When individuals are positively (negatively) misguided, they are more optimistic (pessimistic) about returns to effort than they should be. Note that evaluator assignment treatment manipulates the direction of misguidedness conditional on prior bias. If individuals are initially overconfident (have positive prior bias), they will be positively misguided in MIXED 1, and negatively misguided in MIXED 2. If individuals are initially underconfident (have negative prior bias), they will be negatively misguided in MIXED 1, and positively misguided in MIXED 2. The results of our pre-test data (outlined in Section 4.1) show that participants on average form overconfident performance beliefs ($\bar{\gamma} > .5$), which allows us derive proposition 1.

Proposition 1 *Given average overconfidence in performance beliefs, the average beliefs about the returns to effort among participants in the* MIXED 1 *group are higher compared to participants in the* MIXED 2 *group.*

Given the relationship between performance beliefs γ and inferences regarding returns to effort specified in equations 2 and 3, we derive proposition 2.

Proposition 2 *The difference in average beliefs about the returns to effort between participants in the* MIXED 1 *group and participants in the* MIXED 2 *group is increasing in performance beliefs.*

Causal impact of returns to effort beliefs on actions The experimental variation in the signal generated by evaluator 1 allows us to identify the causal impact of constructed (and potentially misguided) returns to effort beliefs on worker performance in period 2. We propose that the optimal action (effort provision) in period 2 depends on participants' benefits and costs of effort provision. For simplicity, we assume that the output produced by the worker is linearly increasing in the level of effort provision e. Therefore, the returns to providing effort are a linear function of the piece-rate ω and the returns to effort beliefs

θ_1. On the cost side, we assume that effort provision is associated with convex effort costs $c(e) = \frac{1}{2}e^2$.

$$u(e) = \theta_1 \omega e - \frac{1}{2}e^2 \tag{4}$$

Maximizing equation 4 results in the following optimal level of effort provision e^*:

$$e^* = \theta_1 \omega \tag{5}$$

The optimal level of effort provision increases in return to effort beliefs θ_1, which allows us to derive proposition 3.

Proposition 3 *Constructed returns to effort beliefs have a positive and causal effect on effort provision.*

4 Statistical analyses

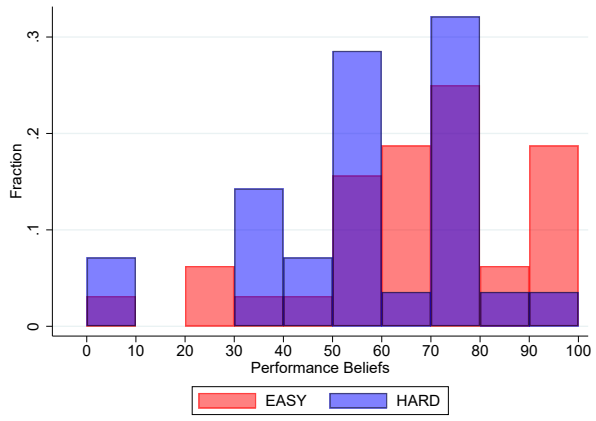
In the analysis, we will use data from participants in the MIXED 1 and MIXED 2 groups who receive mixed payoff information in period 1. We will not analyze data from participants in BOTH HIGH and BOTH LOW groups, since they do not participate in period 2 of the experiment.

4.1 Performance beliefs and manipulation checks

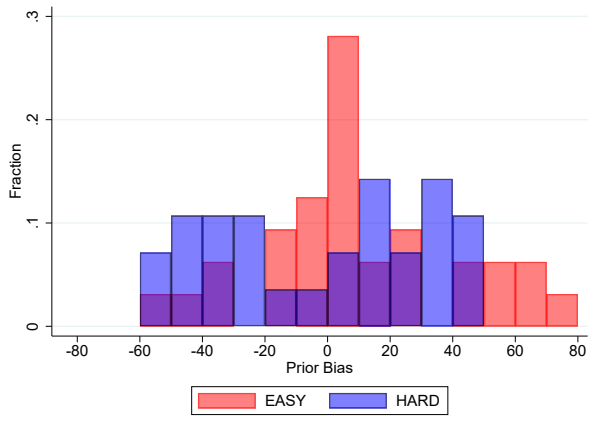
We calibrated the logic quiz such that participants form on average overconfident performance beliefs, especially in the easy quiz version. Panel A in Figure 1 shows the distributions of performance beliefs from our pre-test data, across EASY QUIZ (red bars) and DIFFICULT QUIZ (blue bars) conditions. In Panel B of Figure 1, we plot the distributions of prior bias from our pre-test data, across EASY QUIZ (red bars) and DIFFICULT QUIZ (blue bars) groups. The pre-test data shows that participants in the EASY QUIZ condition hold average performance beliefs of 64.69% while participants in the DIFFICULT QUIZ condition hold average performance beliefs of 53.25%. In line with this result, participants in the EASY QUIZ condition have an average prior bias of 8.98% while participants in the DIFFICULT QUIZ condition have an average prior bias of -1.85%. In the main experiment, we will use the same set of questions for each condition, and in our analyses we will present Figure 1 and use Wilcoxon rank-sum tests to test for statistical difference in performance beliefs and prior bias between EASY QUIZ and DIFFICULT QUIZ conditions.

Figure 1: Manipulation Check

Panel A: Performance Beliefs



Panel B: Prior Bias



4.2 Returns to effort beliefs

Proposition 1 hypothesizes that average beliefs about the returns to effort are higher among participants in the Mixed 1 group compared to participants in the Mixed 2 group when they hold overconfident performance beliefs. We will illustrate the difference of average returns to effort beliefs between Mixed 1 and Mixed 2 groups with bar graphs. In addition, we will test for statistical difference in returns to effort beliefs between Mixed 1 and Mixed 2 groups with a Wilcoxon rank-sum test. Proposition 1 is confirmed when participants in the Mixed 1 group hold significantly higher returns to effort beliefs than participants in the Mixed 2 group.

Proposition 2 hypothesizes that the difference in returns to effort beliefs between the Mixed 1 group and the Mixed 2 group increases in performance beliefs. We will first document group differences in returns to effort beliefs between participants in the Mixed 1 group and participants in the Mixed 2 group across Easy Quiz and Difficult Quiz conditions. We will then estimate a two-stage least squares regression to test for a causal heterogeneous effect. In particular, we will regress returns to effort beliefs on a dummy for being in the Mixed 1 group and the interaction between the dummy for being in the Mixed 1 group and performance beliefs. We will instrument performance beliefs with the quiz difficulty manipulation. Proposition 2 is confirmed when the difference in returns to effort beliefs between the Mixed 1 group and the Mixed 2 group increases in performance beliefs.

4.3 Effort provision

Proposition 3 hypothesizes that returns to effort beliefs have a positive and causal effect on effort provision. Effort provision will be measured by the time spent on the decoding tasks and decoding task performance. We will estimate two-stage least squares regressions to establish the causal impact of returns to effort beliefs on effort provision. In particular, the first-stage will regress returns to effort beliefs on the full-interaction between dummies indicating quiz difficulty and evaluator type conditions (Easy-Mixed 1, Difficult-Mixed 1, Easy-Mixed 2, Difficult-Mixed 2). Proposition 3 is confirmed when the returns to effort beliefs have a significantly positive effect on the time spent on the decoding task and decoding task performance. In addition, we will estimate the difference in group differences in returns to effort beliefs between participants in the Mixed 1 group and participants in the Mixed 2 group across Easy Quiz and Difficult Quiz conditions to test for a causal effect of performance beliefs on effort provision.

4.4 Balance checks and exploratory analysis

We construct an *ability* metric based on participants' performance on questions 1–7 of the logic quiz, which are common across the two quiz versions. We will test for baseline balance in ability between all experimental groups. Recall that we stratified the randomizations of quiz difficulty and evaluator type assignment by observable characteristics. Still, we will test for baseline balance between our quiz difficulty conditions according to gender, and between Mixed 1 and Mixed 2 groups according to gender, ability, quiz difficulty, and performance beliefs. If the sample is unbalanced on a variable that correlates with one of the outcome variables, we will control for the unbalanced correlate in the regressions. In addition, we will explore heterogeneity of our results with respect to gender, ability and deviations from Bayesian updating.

References

Civelli, A., C. Deck, et al. (2018): "A Flexible and Customizable Method for Assessing Cognitive Abilities," *Review of Behavioral Economics*, 5, 123–147.

Danz, D., L. Vesterlund, and A. J. Wilson (2022): "Belief elicitation and behavioral incentive compatibility," *American Economic Review*, 112, 2851–2883.

Heidhues, P., B. Kőszegi, and P. Strack (2018): "Unrealistic expectations and misguided learning," *Econometrica*, 86, 1159–1214.

Hestermann, N. and Y. Le Yaouanq (2021): "Experimentation with self-serving attribution biases," *American Economic Journal: Microeconomics*, 13, 198–237.

Moore, D. A. and P. J. Healy (2008): "The trouble with overconfidence." *Psychological review*, 115, 502.