Hypotheses and planned analyses

Hypothesis 1 (main hypothesis): H0 (Heidhues et al., 2024): In the main treatment, beliefs will converge to a lower task benefit level than the true one: $\hat{\theta}4 < 120$; H1: $\hat{\theta}4 = 120$. If we do not find that people are able to update correctly in the control treatment (i.e., $\hat{\theta}4$ is statistically different from 120 points), we will use the control treatment beliefs (instead of 120) as benchmark for the main treatment analysis (i.e., H0: participants will update upwards less in the main than in the control treatment).

Analyses: (i) summary statistics/visualization how the beliefs change from before the trial round to part 5 for both treatments, (ii) test whether beliefs in the main treatment are statistically different from 120 points in parts 5 (or the corresponding control treatment beliefs), using the Wilcoxon signed-rank test (one sample median test) and a one sample t-test (or alternatively, Mann-Whitney-U test and two-sided t-test to test against control treatment beliefs).

Hypothesis 2: In the control treatment, the belief about the fundamental will converge to the true level over time.

Analyses: (i) summary statistics/visualization how the belief changes from before the trial round to part 5, expecting the belief to be larger in part 5 than in part 4 than in part 3 than before the trial and in part 2, (ii) test whether the belief is statistically different from 120 points in part 5, using the Wilcoxon signed-rank test (one sample median test) and a one sample t-test.

Hypothesis 3: The number of correctly solved encryption tasks will be higher in the control treatment with the known task benefit than in the main treatment with uncertain task benefit and an initially lower belief on the task benefit in all periods. Caveat on maximum effort provision in real effort tasks applies.

Analyses: (i) summary statistics/visualization how the number of solved encryption tasks changes from across all parts in both treatments, (ii) MWU-tests testing for differences in the number of correctly solved encryption tasks (averaged over parts and by part) across treatments, (iii) panel data regression of number of correctly solved encryption tasks on a treatment dummy and an interaction between treatment and part (to investigate a possible time trend, explorative)

Hypothesis 4: In both treatments, predictions on the number of correctly solved encryption tasks will converge to the actual number of correctly solved encryption tasks in part 5, as (if) participants become sophisticated.

Analyses: (i) test if (prediction – actual) is statistically different from 0 in part 5 using the Wilcoxon signed-rank test (one sample median test) and a one sample t-test, (ii) test if (prediction – actual) is smaller in part 5 than in earlier parts using the Wilcoxon signed-rank test and a t-test.

Hypothesis 5: Lower beliefs on task benefit levels lead to lower stated numbers of ideally solved encryption tasks (within individual). H0 (Heidhues et al., 2024): The stated number of ideally solved encryption tasks will be lower in part 5 in the main treatment than in the control treatment. H1: The stated number of ideally solved encryption tasks will be as high in part 5 in the main treatment as in the control treatment.

Analyses: (i) summary statistics/visualization of the ideal number of solved encryption tasks in part 5 in the main treatment and the control treatment, (ii) Mann-Whitney U tests and two-sample t-test for the difference in the ideal number of solved encryption tasks in part 5 in the main treatment and in the control treatment.

Further aspects

While self-control problems (present-bias) are pervasive (Cobb-Clark et al., 2024), about 25% of people in population-representative data from Germany do not have any self-control problems. Furthermore, some people may be future-biased instead of present-biased or overly pessimistic regarding their own self-control problems. However, the hypotheses from Heidhues et al. (2024) above refer to people who do have self-control problems, i.e., are present-biased and not fully sophisticated, underestimating their self-control problems. We will therefore run two kinds of analyses: first, one using the data from all participants to learn whether the predictions of Heidhues et al. (2024) provide an account of average behavior; second, we will run the same kind of analyses restricting our sample to present-biased, not fully sophisticated individuals that Heidhues et al. (2024) refer to.

For the second kind of analysis, we will classify participants in the following way: In the absence of learning about task benefits and costs, ideal, predicted, and actual future choices coincide for time-consistent individuals (O'Donoghue and Rabin, 1999). We will elicit each participant's ideal and predicted number of correctly solved encryption tasks for given task benefits of at the beginning of part 1, after they have gained experience on working on the task in the trial round such that participants are aware of their productivity in the task and how costly the task is for them and before they will actually work on the encryption task again. We will classify (close to) time-consistent individuals as those for whom the ideal, the predicted, and actual number of correctly solved encryption tasks in part 1 are the same or only differ by a margin of +/- five percent around the actual number of correctly solved encryption tasks. For that purpose, in the main treatment, we will condition the ideal and predicted number of tasks on an individual's prior belief on task benefits and extrapolate if necessary. We will apply the same margin around actually solved tasks to classify fully sophisticated individuals (for whom predicted = actual < ideal). Individuals who are overly pessimistic regarding their own self-control problems are characterized by an actual number of solved tasks that is larger than the one stated as predicted.

Cobb-Clark, D. A., Dahmann, S. C., Kamhöfer, D. A. & Schildberg-Hörisch, H. (2024). Sophistication about self-control. *Journal of Public Economics*, *238*, 105196.

A central goal of our study is to identify whether participants update their beliefs about the task benefit rationally or misinterpret themselves based on their past behavior.

To test for forgetting, we include a short **recall task** at the end of the experiment. Participants are asked whether they remember specific numbers from the signals they observed (e.g., how often they saw a particular number) or whether they cannot recall them. This measure allows us to assess the extent to which participants remember or forget information about the task benefit. Forgetting of these signals would support the interpretation that belief updating reflects self-justification rather than rational information processing. In addition, we ask whether they have taken a screenshot of the signals during the experiment, as this indicates whether they tried to keep a record of the information they received.

To capture heterogeneity in the cognitive demands of belief updating, we include two additional measures. First, we administer an **average calculation task**, in which participants are asked to compute the mean of five given numbers. This provides a simple check of participants' numerical understanding, which is directly relevant for processing the signals and updating beliefs. Second, we use the **Digit Span subtest** of the Wechsler Intelligence Scale to measure working memory capacity, which may influence the ability to store and integrate information across parts.

The hypotheses above serve the purpose to investigate whether "rational updating" or "misinterpreting oneself" explain the average observed updating behavior better, which is the focus of our analysis. On top of that, we can try to identify those types at the individual level. In principle, rational updaters should (roughly) update their prior according to Bayes' rule given the signals they have seen. Individuals who misinterpret themselves should update to lower levels as implied by Bayes' rule. However, cognitive constraints (e.g., limited memory or low mathematical skills) may hinder people from updating according to Bayes' rule. The digit span test and the average calculation task allow for classifying people according to their capability for rational updating. Only those who correctly solve the average calculation task and perform high in the digit span test will qualify for such a classification at the individual level.

Primarily for payment purposes, we record the time participants spend on the encryption task versus on alternative, non-work activities. Participants can freely allocate their time between completing encryption tasks, which generate higher earnings, and non-work alternatives, which provide a small flat fee. The non-work activities are included as an alternative to create a setting where self-control problems may arise. While the measurement of time allocation is mainly used for payment purposes, it also provides an additional measure of work effort. For example, the share of time spent on the encryption task minus time on encryption, divided by total time – can be analyzed alongside the main effort measure, the number of correctly solved encryption tasks.

To generate time lags between different parts of the experiment and foster forgetting of previously seen signals, participants additionally answer survey questions eliciting sociodemographics, economic preferences, the Big Five personality traits, and the Brief Self-Control Scale (BSCS) by Tangney et al. (2004). These questions primarily serve as fillers to create a delay between experimental parts. We may use them to study heterogeneity of our main hypotheses in these dimensions.

We measure participants' risk aversion using a lottery choice list. We will split our sample into risk-averse versus non-risk-averse individuals and check whether their outcome variables differ.