

case 09 xiao progress

Participant-visible method context. Results, discussion, captions, and the target figure are excluded.

The study consisted of two online experiments. Experiment 1 involved collecting a discovery sample for exploratory analysis, while Experiment 2 was designed as a replication and extension of Experiment 1. Experiment 1 was not preregistered. Experiment 2, including methods and hypotheses, was preregistered at aspredicted.org (<https://aspredicted.org/4t26-7wg9.pdf>).

Participants For Experiment 1, we recruited 30 participants, with the sample size determined by heuristics common in experimental psychology. For Experiment 2, to obtain a power of 95% to detect a correlation of $r = .3$ for our hypothesis on interindividual differences (i.e., for the correlation between emotional volatility and the preference for progress feedback, using a one-tailed test as per the preregistration), we aimed for $n = 120$ postexclusion. This sample size would also be sufficient to achieve 95% power for all other preregistered hypotheses, when we based the effect size estimate on results from Experiment 1 and a previous study on progress feedback (see S2: Power Analysis in the online supplemental materials for further details). We therefore

VALUE OF PROGRESS FEEDBACK IN PHYSICAL EFFORT

recruited an initial sample of 137 participants online via Prolific (Palan & Schitter, 2018). The study was approved by the Tel Aviv University local ethics committee (Reference: IRB 0003132-2; IRB refers to the Institutional Review Board), and all participants provided (anonymized) informed consent before participation. For both experiments, every participant was given a base payment of approximately £8.00 per hour, with a bonus payment of up to £1.2 based on task performance. Both studies used the inclusion criteria of ages 18–80, no ongoing mental health conditions, and no neurological conditions (e.g., dementia, mild cognitive impairment, or multiple sclerosis). As an additional exclusion criterion, we asked participants with an implausibly low maximal pulling capacity (see below) of less than 1 pull per second to return the study submission, and they could not start the main task. In Experiment 1, two out of 32 participants were screened out with this method, leaving 30 participants (rangeage = 23–68, $M = 43$ years, $SD = 12$, 10 female). In Experiment 2, two out of 139 participants failed this test, leaving a total of 137 participants who completed the task (rangeage = 18–72, $M = 35$ years, $SD = 11$, 34% female). All 30 participants were included in the analyses for Experiment 1. For Experiment 2, among the 137 participants, two participants had incomplete data because of a technical error and could not be analyzed; 15 participants were excluded from the main analyses based on the preregistered exclusion criteria. Six more participants were excluded from the main analysis because of an unexpected issue that resulted in a deviation from the preregistration: these participants exhibited implausible values that would bias the parametric statistical analysis; we conducted sensitivity analysis on the preregistered hypothesis to show that this exclusion did not affect the main findings. The remaining $n = 114$ (rangeage = 18–72, $M = 36$ years, $SD = 12$, 34% female) retained 95% power to detect the preregistered effect size of interest under the preregistered one-tailed tests. Details about the exclusion criteria and quality control measures are reported in full in S1: Supplementary Methods in the online supplemental materials.

Ore Mining Task We developed a new online physical effort task, the “Ore Mining Task,” designed to measure continuous movements and button clicks (Figure 1). This gamified task operationalizes effort through the alternating clicking of two vertically offset buttons using a mouse or a trackpad. Participants engage with a virtual “pulley” system to simulate pulling ore from deep underground, which needs to be completed within a time limit, after which the ore would “oxidize” and lose its value. In Experiment 1, the time limit was 8 s. In Experiment 2, the time limit was increased to 15 s in order to enrich the continuous movement data provided by cursor tracking. To lift the ore, participants alternately clicked on the two buttons, with each set of clicks representing a single “pull.” The pulley was deliberately depicted as damaged, shown by a zigzagged wheel, meaning that not every pull would lead to moving the ore upward. The depth of the ore, marked by a red line, indicated the effort level associated with each option. Importantly, a lower red line indicated that more work (i.e., pulls) was required, corresponding to a higher effort level (gamified as a greater depth from which the ore had to be pulled). Comprehension of this fundamental aspect of the task was tested through a quiz question, and participants were required to answer it correctly before beginning the main task. The number

of required pulls at each effort level was individually calibrated based on each participant’s maximum pulling capacity (MPC) within the time limit (i.e., 8 s for Experiment 1 and 15 s for Experiment 2) obtained at the start of the

task (S1: Supplementary Methods in the online supplemental materials). The task had three phases: decision phase, effort exertion phase (“pulling”), and effort rating phase. In the decision phase, participants chose between two stations: with progress feedback (PF+) and no progress feedback (PF-). In PF+, participants could observe the ore moving up toward the surface with their pulling, while in PF-, the ore mining station was hidden behind a black rectangle, occluding visual feedback on current progress. The effort levels offered in PF- were adjusted through a fixed one-up/one-down staircase procedure. Specifically, in Experiment 1, choosing the PF- resulted in a 10% MPC increase in the PF- effort level and choosing the PF+ decreased the effort level for PF- by 10% for a given staircase. In Experiment 2, to improve the sensitivity to subtle differences in the preference for PF+, the step size was reduced from 10% to 5%. Thus, the effort level associated with PF- was dynamically adjusted based on participants’ choices. Overall, if a participant consistently chose PF-, the effort level for PF- increased on subsequent trials relative to PF+. This continued until the participant switched to choosing PF+. In this way, the procedure adaptively identified the individualized point at which a participant was indifferent between receiving progress feedback and exerting less effort. Interleaved staircases were used to conceal the nature of adaptation. Within each staircase, the effort level of PF- changed based on the procedure above, while the PF+ effort level was constant. In Experiment 1, four staircases were interleaved (PF+ effort levels fixed at 70%, 80%, 90%, or 100% of each participant’s MPC), while in Experiment 2, this was reduced to three staircases (PF+ fixed at 70%, 85%, or 100% of the MPC) in order to reduce the overall trial number. After selecting one of the two stations, the chosen station appeared at the center of the screen, which indicated the start of the effort exertion phase. To introduce uncertainty in progress, not all pulls resulted in the ore moving upward. The number of “stuck” pulls was fixed at 30% of each participant’s MPC for all trials in both experiments, and these events were randomly distributed across the required pulls. If the required number of pulls was completed within the trial time limit, the buttons turned green, and participants gained a fixed reward of 10 points. If time was up before completing the required number of pulls, the buttons turned red, and participants did not gain any points. Feedback on whether the ore was successfully lifted was provided on all trials, so that the feedback about performance was dissociated from progress feedback. Participants were told that at the end of the game, they would receive an extra bonus payment based on the number of points they earned. Finally, after completing the effort exertion, participants entered the effort rating phase. In the effort rating phase, after completing the effort exertion phase and receiving feedback on performance, participants were asked to rate how effortful the last trial felt. For Experiment 1, the task involved 60 trials. Experiment 2 was shortened to 45 trials, which was able to retain adequate parameter recovery for the computational model (see S3: Details on Computational Modeling in the online supplemental materials). Participants familiarized themselves with exerting effort under PF- and PF+ through six practice trials. To mitigate fatigue, we reminded participants that they could rest on every choice screen. In Experiment 2, after the halfway point, we recommended a further

XIAO AND WOLPE

Figure 1 Schematics of a Single Trial in the Ore Mining Task

Note. PF+ = with progress feedback; PF- = no progress feedback. See the online article for the color version of this figure.

rest of at least 1 min. The experiment was written in jsPsych 7.3 (de Leeuw et al., 2023), with custom-written plugins for the effort exertion phase, and ran on a server with Just Another Tool for Online Studies (Lange et al., 2015) backend.

Questionnaires In Experiment 2, which was powered to investigate interindividual differences, we included the following self-report questionnaires: the apathy motivation index, which measures three facets of apathy (Ang et al., 2017); the Patient Health Questionnaire for symptoms of depression (Kroenke et al., 2001); the intolerance of uncertainty scale, as an additional measure of aversion to the uncertainty that accompanies the absence of progress feedback (Carleton et al., 2007); and 12 items from the energy level, anxiety, and emotional volatility facets of the next Big Five Inventory (BFI-2; Soto & John, 2017). The BFI-2 was chosen because openly available data in a previous study (Bennett et al., 2021) suggested that the four-item emotional volatility subscale had the strongest positive correlation with the cost a person is willing to pay for information; we tested whether this relationship extends to the willingness to exchange effort cost with progress information.

Computational Model of Choice To test whether participants preferred progress feedback, we fitted variants of logit choice models to each participant’s decisions to estimate the subjective value of progress feedback (SVPF+):

$$p(\text{choose PF+}) =$$

$$1 - \text{lapse} \cdot \frac{1}{1 + e^{-\beta(\Delta\text{effort} - \text{SVPF+})}}$$

(1)

This model assumes that choices are driven by the difference in effort levels between PF+ and PF− (Δeffort), as well as the individual SVPF+, which is the subjective value a person assigns to progress feedback (equivalent to the point of subjective equivalence in psychometric models; a positive subjective value can offset Δeffort to make PF+ more preferable). The inverse temperature, β , controls the slope of this relationship and is expected to be negative. The more negative β is, then for each unit increase in Δeffort , the probability of choosing PF+ decreases. The lapse parameter controls the asymptote of the decision function and aims to capture a fixed rate of random choices that occur irrespective of the effort difference. A high lapse rate would indicate random and inattentive responding, while a lapse rate of zero would mean that, when the Δeffort is sufficiently large, the participant will always choose the low-effort option. We compared models where the lapse rate was fixed to zero or allowed to vary and where SVPF+ was fixed to zero or allowed to vary. A model with SVPF+ fixed to zero corresponds to the hypothesis that all participants preferred the option with the lower effort level, with no additional preference for progress information. As a control model, we also included a choice kernel model, where participants have a fixed tendency to select the previously chosen option with no consideration for the effort levels (Wilson & Collins, 2019). The models were fitted with maximum likelihood estimation using `fmincon` in MATLAB R2020a (MathWorks, Natick, Massachusetts, United States), and the Bayesian Information Criteria (BIC) were calculated for each participant as an approximation of model evidence. Model comparison was performed on the mean BIC across participants. For the winning model, a one-sample t test was used to test whether SVPF+ across participants was significantly greater than zero.

General Statistical Analysis For all mixed-effects models, we used the `lme4` package (Bates et al., 2015), as well as `lmerTest` for Satterthwaite's method of approximating the degrees of freedom (Kuznetsova et al., 2017). The analysis was run in R Version 4.2.3. We fitted linear mixed-effects models to the trial-by-trial effort rating data. Fixed effects were the effort levels (%MPC) and effort type (PF+ vs. PF−), as well as alternative models with or without corresponding random slopes. We used BIC to select the best-fitting model.

VALUE OF PROGRESS FEEDBACK IN PHYSICAL EFFORT

As the task involved a staircase procedure, participants experienced different trial numbers for the different conditions. Moreover, participants occasionally failed to complete the effortful task. Therefore, for Experiment 2, as preregistered, a more stringent exclusion criteria were applied to ensure that the results were not driven by these instances. Thus, only participants who have experienced a range of conditions and have multiple repeated measurements within these conditions were included in the confirmatory analysis (see section 6 in the additional online materials at <https://aspredicted.org/4t267wg9.pdf>), with $n = 88$. Importantly, as per preregistration, we also conducted a sensitivity analysis without the additional criteria. An α level of .05 was used for all null hypothesis significance tests, while for BIC, a difference in BIC ≥ 6 constitutes strong evidence for a model, while a difference of 2–6 constitutes positive evidence. For Experiment 2, we preregistered one-tailed tests for the hypothesized effects; otherwise, for Experiment 1 and for all exploratory analyses in Experiment 2, two-tailed tests were used.

Analysis of Continuous Cursor Movement In Experiment 2, during effort exertion, we recorded cursor movement data. These are the timestamped positions of the cursor as it moved between the two buttons during effort exertion, which reflect real-world hand or finger movement with a mouse or a trackpad, respectively. To analyze continuous movement trajectories, we preprocessed the position and velocity data recorded during the task. For each participant and trial, the raw data included time-stamped x and y coordinates of the cursor movement. Cursor velocities in the horizontal (x) and vertical (y) directions were computed as the first derivative of position with respect to time. To reduce high-frequency noise occurring because of occasional device sampling errors, the resulting velocity signals were smoothed using a moving mean filter. As the raw sampling frequency was approximately 75 Hz for most devices, we chose a window size of six samples, which effectively minimized noise while preserving the integrity of the signal. Outlier velocities in the vertical direction were identified using an absolute z score threshold of 3 and replaced via linear interpolation. The cleaned velocity signals were then used to derive trial-by-trial movement features. Specifically, we extracted the maximum velocity and variance of velocities in both the x and y directions, as well as the cumulative path length, computed as the sum of Euclidean distances between consecutive positions. Path variability was quantified as the variance in normalized x and y positions, scaled relative to the task window size.

