PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Wagers for work: Decomposing the costs of cognitive effort

Sarah L. Master[1]*, **Clayton E. Curtis**[1,2], **Peter Dayan**[3,4]

1 Department of Psychology, New York University, New York, New York, United States of America, **2** Center for Neural Science, New York University, New York, New York, United States of America, **3** Max Planck Institute for Biological Cybernetics, Tübingen, Deutschland, **4** University of Tübingen, Tübingen, Deutschland

* sm4937@nyu.edu

OPEN ACCESS

Citation: Master SL, Curtis CE, Dayan P (2024) Wagers for work: Decomposing the costs of cognitive effort. PLoS Comput Biol 20(4): e1012060. https://doi.org/10.1371/journal.pcbi.1012060

Editor: Wouter Kool, Harvard University, UNITED STATES

Received: July 31, 2023

Accepted: April 10, 2024

Published: April 29, 2024

Data Availability Statement: All task, computational modeling, and analysis code relevant to this project is freely available on github: https://github.com/sm4937/coce The full dataset is publicly available on the Open Science Framework: https://osf.io/8hz6v/.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Some aspects of cognition are more taxing than others. Accordingly, many people will avoid cognitively demanding tasks in favor of simpler alternatives. Which components of these tasks are costly, and how much, remains unknown. Here, we use a novel task design in which subjects request wages for completing cognitive tasks and a computational modeling procedure that decomposes their wages into the costs driving them. Using working memory as a test case, our approach revealed that gating new information into memory and protecting against interference are costly. Critically, other factors, like memory load, appeared less costly. Other key factors which may drive effort costs, such as error avoidance, had minimal influence on wage requests. Our approach is sensitive to individual differences, and could be used in psychiatric populations to understand the true underlying nature of apparent cognitive deficits.

Author summary

Anyone who has tried to mentally calculate how much to tip at a restaurant knows that cognitive effort can feel aversive. Doing math in your head, like most high-level cognitive abilities, depends critically on working memory (WM). We know that WM is sometimes effortful to use, but we don't know which aspects of WM use drive these effort costs. To address this question, we had participants request wages in exchange for performing various tasks that differed in their specific WM demands. Using computational models of their wage demands, we demonstrated that some aspects of WM are costly, such as bringing new information into memory and preventing interference. Other factors, like the amount of information in memory and attempts to avoid mistakes, were less costly. Our approach identified which specific subcomponents of WM are aversive. Future research could use these methods to test theories about how motivational problems might be masquerading as cognitive deficits in psychiatric populations.

Introduction

Some activities (e.g., getting dinner with friends) are more enjoyable than others (e.g., calculating how to split the bill). Doing tasks which require greater cognitive effort, colloquially called "brain power," can feel uniquely aversive, though to different degrees for different people [1–3]. Indeed, despite tangible benefits, people often avoid cognitively demanding work [4,5]. Such resistance suggests that we weigh the effort of mental activity, perhaps as a cost to be offset with reward.

Previous research has identified the experimental tasks which are more costly to perform by giving subjects control over which tasks they complete. Tasks which subjects demanded the most incentives to complete [5–7] or which subjects tended to avoid in favor of other tasks with equivalent rewards [4,8,9] are considered most effortful. Some costly aspects of these tasks are external, like time on task [10–12] or the complexity of the cognitive model required by the task [13–16], but other costs arise from the internal operations necessary to realize external actions, like the degree of hierarchical cognitive control required [17]. In general, cognitive resistance increases when tasks place substantial demands on working memory and cognitive control [5,18–20]. However, it remains unknown which particular aspects of working memory and cognitive control may be most costly. For example, perhaps the sustained effort required during working memory maintenance is more costly than the transient effort required to inhibit a prepotent response.

Here, we decomposed simple and complex attention (i.e., detection) and working memory (N-back) tasks into putative elemental processes such as maintaining information in memory of different loads and resisting interference from task irrelevant lures. We assumed that the subjective costs of these operations are internally felt and consciously accessible, and that the total cost of completing a task is learned by experiencing these costs. We assessed these total costs using a modified auction procedure. Previous work has used such auctions to infer the subjective values of items on a menu [21,22]; our modifications allowed us to infer the total effort costs associated with completing various cognitive tasks by asking subjects what a "fair wage" for task completion would be. Given the evidence that the allocation of cognitive resources is subject to a cost-benefit tradeoff [12,23–27], we hypothesized that subjects' trial-by-trial fair wage demands would, at least to a first approximation, reflect the sum of the individual costs associated with task completion, as the amount of reward necessary to offset them. To assess the extent to which the costs we measured were related to the self-reported tendency to engage in effortful cognitive tasks, albeit varying across trials, we collected Need For Cognition scores from each of our subjects (NFC) [2].

As our subjects were likely to experience costs other than those deriving from cognitive effort, we designed our experiment to try to limit the effects of these other factors. First, to minimize the influence of time on task on fair wage ratings, we gave subjects an easy task to complete when they wished to skip a harder one. We also ensured that every task round took the same amount of time. Second, cognitively effortful tasks often also elicit errors. This may be experienced as a cost, particularly in perfectionist subjects [28,29]. While we could not completely control for error avoidance costs as for time costs, we designed our task to minimize error avoidance behavior by not giving trial-by-trial feedback, not informing subjects of their accuracy round-by-round, and not reducing their compensation unless errors became overly prevalent. We also collected subject scores on the Short Almost Perfect Scale (SAPS) [30], to assess the degree to which subjects' fair wages were driven by the tendency to avoid making errors (i.e. perfectionism). Lastly, we included the costliness of errors alongside the costs of cognition in our computational analyses.

We found three non-zero cognitive effort costs: the cost of adding new information into working memory (WM), the cost of filtering out irrelevant information, and the cost of maintenance. More subtly, we found evidence that subjects learned the total costs of each task through task experience, and that the costs of cognition did not increase or decrease over the duration of the experiment. We found that the self-reported tendency to avoid effort was related to explicit ratings of task costliness and difficulty, as well as more implicit costs of cognition. This implies that effort avoidance may be driven both by the explicit, stable preference to avoid effort and by the implicit subjective experiences of the costs of cognition.

Results

100 subjects completed the experiment online through Amazon Mechanical Turk. Subjects completed 32 task rounds and performed four different tasks in random order: an attentional vigilance task (1-detect), a vigilance task requiring more WM maintenance (3-detect), and the 1- and 2-back WM task [19,31]. Every task involved monitoring the screen in order to make button presses in response to semi-infrequent target stimuli. In the 1-detect task, subjects were simply in search of the target letter "T". In the 3-detect task, the target was any letter presented thrice in a row (e.g. three R's in sequence). In the N-back tasks, subjects memorized the incoming sequence of letters in order to identify when the letter on screen matched the letter presented N (1 or 2) trials ago. Before each task round began, subjects were shown the task they were to complete (an associated fractal), and were able to request a fair wage for that round of that task. A Becker-Degroot-Marschak (BDM) auction mechanism then determined whether they completed 15 trials of the task they rated for the wage they requested, or 15 trials of the default, non-demanding task (the 1-detect) for a lower wage (Fig 1). We analyzed their performance and fair wage ratings across tasks. We used computational modeling to examine how fair wages were influenced by the putative cognitive operations used to complete the previous task rounds, like WM maintenance or updating. We also related fair wage ratings to previous task behavior, including the number and types of errors they made.

Model-Agnostic results

There was a main effect of task identity on accuracy (Fig 2A; $F = 44$; $p < 0.001$), mean reaction time (RT; $F = 31$, $p < 0.001$), and difficulty rating ($F = 26$; $p < 0.001$). Post-hoc comparisons confirmed that subjects had lower accuracy and higher RTs on the 2-back task than on all of the other tasks (Table 1; Accuracy: 2-back versus 1-detect $t = 12$, $p < 0.001$; 2-back versus 1-back $t = -6.0$, $p < 0.001$; 2-back versus 3-detect $t = 9.6$, $p < 0.001$; Mean RT: 2-back versus 1-detect $t = -12$, $p < 0.001$; 2-back versus 1-back $t = 7.3$, $p < 0.001$; 2-back versus 3-detect $t = -14$, $p < 0.001$). Subjects also rated the 2-back as more difficult than the 1-detect ($t = -4.7$, $p < 0.001$) and the 3-detect ($t = 5.8$, $p < 0.001$). They rated the 1- and 2-back as equally difficult ($t = -0.40$, $p = 0.69$). Accuracy was highest on the 1-detect when compared with all the other tasks (vs 1-back $t = 6.3$, $p < 0.001$; vs. 3-detect $t = 5.4$, $p < 0.001$; vs. 2-back $t = 12$, $p < 0.001$). Mean RTs on the 1-detect were lower than on the 1-back ($t = -7.4$, $p < 0.001$), and 2-back ($t = -13$, $p < 0.001$), but not on the 3-detect ($t = 1.8$, $p = 0.07$). Difficulty ratings were also lowest on the 1-detect compared to the 1-back ($t = -5.0$, $p < 0.001$), 3-detect ($t = -8.5$, $p < 0.001$), and 2-back ($t = -4.6$, $p < 0.001$). Mean accuracy was lower ($t = 3.9$ $p < 0.001$) and mean RT was higher on the 1-back than on the 3-detect ($t = -7.6$, $p < 0.001$). The mean difficulty rating was no different between the 1-back and 3-detect ($t = 6.4$, $p = 0.69$).

The task difficulty ratings collected at the conclusion of the experiment speak to how subjects perceived their own performance on these tasks. In general, subjects' task difficulty ratings were highly correlated with their mean accuracy on that task (1-detect accuracy versus

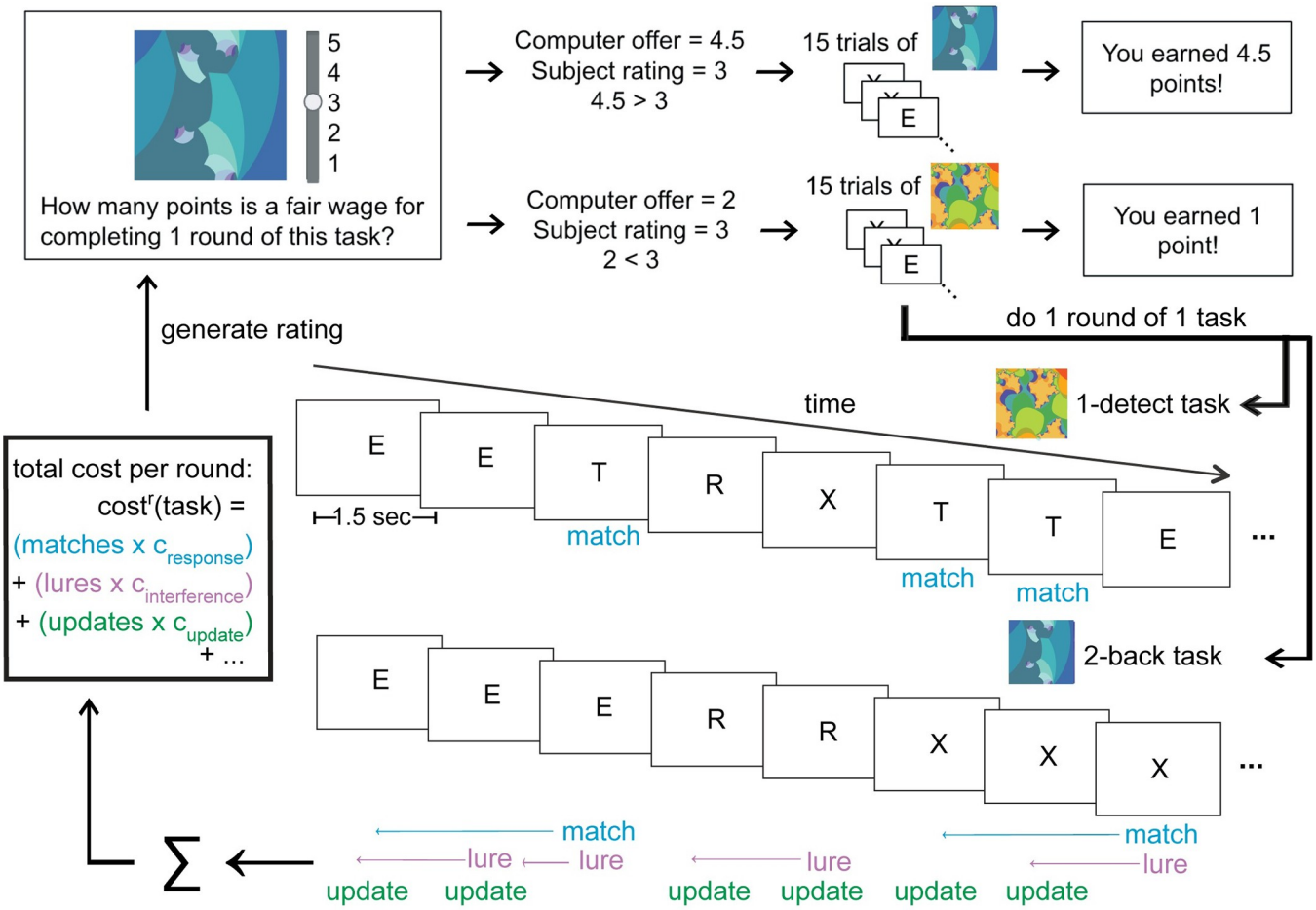

Fig 1. The behavioral paradigm & computational modeling approach. Before each round of the experiment, subjects were shown an image which was associated with one of three possible tasks. They then indicated the wages (in points) that they would like to receive for completing 1 round of that task. If their fair wage rating was below a random computer offer, then they would complete that task and receive the computer's offer. If their fair wage was above a random computer offer, then they would complete a different, easier task instead. We employed this inversion of the Becker-Degroot-Marschak auction procedure to incentivize subjects to be truthful in their fair wage ratings. The fractal images in this figure were obtained under a Creative Commons Zero 1.0 Public Domain License (https://creativecommons.org/publicdomain/zero/1.0/) from openclipart.org (https://openclipart.org/detail/300064/colorful-abstract-background and https://openclipart.org/detail/310263/another-abstract-background).

https://doi.org/10.1371/journal.pcbi.1012060.g001

difficulty rating r = -0.44, p <0.001; 1-back, r = -0.32, p = 0.002; 3-detect, r = -0.37, p <0.001), with the exception of the 2-back task (relationship between accuracy and difficulty rating r = -0.0030, p = 0.98). This indicates that while perceived difficulty and actual error rates were related for most of the tasks, subjects felt the 2-back task was more difficult than it was.

Subjects provided fair wage ratings for the 1-back, 2-back, and 3-detect tasks. As the 1-detect task was the default task, we did not obtain fair wage ratings on it, and assumed subjects would use it to calibrate their other fair wage ratings, as it was worth a fixed amount of points (1 point per round). A 2-way ANOVA on fair wage ratings showed a main effect of task identity (Fig 2B and Table 1; F = 29, p <0.001) and a main effect of task iteration (Figs 2D and S1; F = 5.2, p <0.001). Subjects' mean fair wage ratings on the 2-back task were significantly higher than for the 1-back (t = -8.6, p <0.001). Comparing fair wage ratings for the 1- and 2-back allows us to directly measure the costs of maintaining one more item in working memory, though the 1- and 2-back tasks also differ in the degree of interference present in WM and the number of errors made. Mean fair wage ratings on the 2-back were also higher than on the

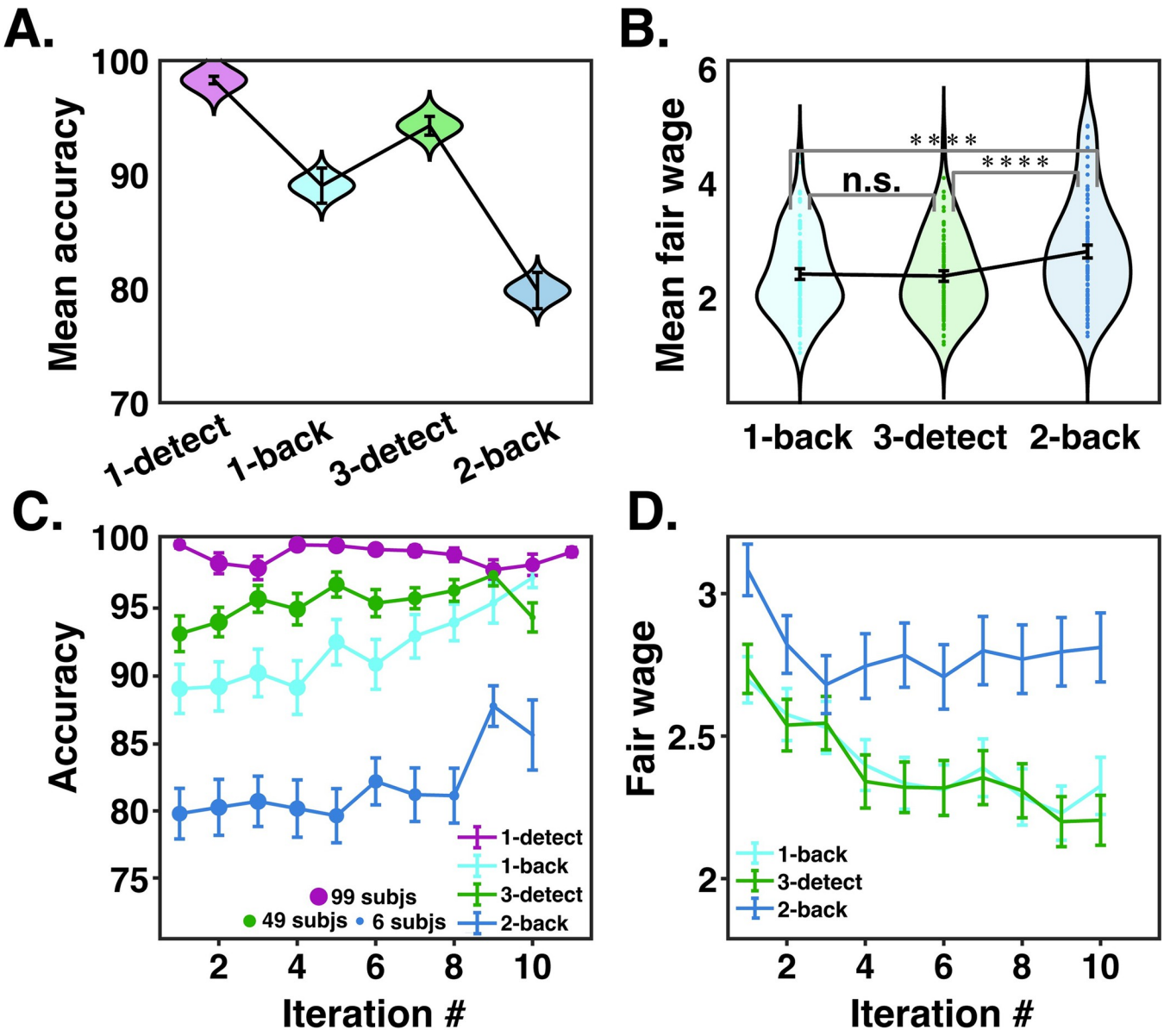

Fig 2. Model-agnostic behavioral results. A. Distributions of mean accuracies across all subjects for the default task (1-detect), and the three rated tasks (1-back, 3-detect, and 2-back). The black bars depict the means and standard errors of the mean (SEMs) of each distribution. The distribution of all subjects' mean accuracies was plotted using a Gaussian kernel via violin.m. All mean accuracies for each task were significantly different from each other (all p's < 0.001). **B.** Distributions of mean fair wages across all subjects for the three rated tasks. The lowest possible rating was 1, and the highest possible rating was 5. The black bars depict the means and SEMs of each distribution. The distribution of ratings was plotted using violin.m. **** indicates significance at the $p < 0.0001$ level. **C.** Mean accuracy across all subjects on each iteration of each task. Due to the stochasticity inherent to the BDM auction procedure, individual subjects completed the 1-back, 3-detect, and 2-back tasks a variable number of times, but a maximum of 11 times each. The relative number of subjects who completed each iteration is depicted by the size of the dot plotted at the mean. Error bars are SEMs. A two-way ANOVA of task and task iteration revealed a main effect of task identity ($F = 15$, $p < 0.0001$) but no effect of task iteration ($F = 1.3$, $p > 0.05$). Thus mean accuracy was different across tasks but did not change with task experience. **D.** Mean fair wage rating by rating number, where the maximum is 11 ratings of one task. A 2-way ANOVA on BDM ratings showed a main effect of task identity (Table 1; $F = 33$; $p < 0.0001$) and a main effect of task iteration (Fig 1; $F = 21$; $p < 0.0001$). Error bars are SEMs.

https://doi.org/10.1371/journal.pcbi.1012060.g002

3-detect ($t = -9.7$, $p < 0.001$). Comparing fair wages from the 2-back and 3-detect, which both require the maintenance of 2 items, allows us to measure the cost of interfering stimuli in WM storage or the increased errors made on the 2-back task. Fair wages were not significantly different between the 1-back and the 3-detect ($t = 0.85$, $p = 0.39$). Though the 1-back and 3-detect

Table 1. Mean accuracy, reaction time (RT) in milliseconds, and difficulty ratings across all subjects for the default task, the 1-detect, and for the three rated tasks, the 1-back, 3-detect, and 2-back tasks. The maximum RT was 1500 milliseconds. The minimum fair wage and difficulty rating was a 1, and the maximum was a 5.

Group means	1-detect	1-back	3-detect	2-back
Percent accuracy	98	89	94	80
RT (msec)	550	611	530	720
Difficulty rating	1.9	2.4	2.4	3.3
Fair wage	NA	2.4	2.4	2.8

https://doi.org/10.1371/journal.pcbi.1012060.t001

differ in their load on WM, subjects tended to rate them equivalently. These results suggest that increasing WM interference may be more subjectively costly than increasing WM load. We investigate further in our model-based analyses below.

While accuracy was significantly lower and fair wages significantly higher for the 2-back task, there was no relationship between mean 2-back accuracy and mean fair wage on the 2-back across subjects ($r = -0.09$, $p = 0.41$). There was also no relationship between mean accuracy and fair wage on the 3-detect task ($r = -0.17$, $p = 0.11$). However, there was a significant relationship of 1-back mean accuracy and mean fair wage ($r = -0.36$, $p < 0.001$). We further assess the influence of errors on fair wages in the model-based analyses below.

Task accuracy was broadly stable across task iterations (Fig 2C; main effect of task iteration $F = 1.3$, $p = 0.23$). This indicates that performance did not improve with task experience. Across all subjects there was also no relationship between round number (out of 32) and mean task accuracy (Pearson $r = -0.02$, $p = 0.14$); or mean RT (Pearson $r = 0.0034$, $p = 0.85$). This is likely because subjects trained to 80% accuracy during practice and were already at their maximum performance levels by the start of the main task.

Fair wage ratings seem to decrease with task iteration ($F = 5.2$, $p < 0.001$; Fig 2D), but potentially as a byproduct of the experimental design. That is, subjects who asked for lower wages completed the non-default tasks a higher number of times; therefore the lower mean fair wage on later task iterations may primarily come from subjects who had lower fair wage ratings overall (S2 Fig). Another possibility is that subjects ask for lower fair wages over time because they find that the tasks become less effortful with practice. If that were the case, then you might expect their accuracy to improve over the course of the experiment. However, the ANOVA on task accuracy by task iteration reported above found no main effect of task iteration. We investigated this further by averaging fair wages over each subject's first and last half of task completions, and comparing them via t-test to see whether their wage requests changed as their task experience increased. We did the same analysis for task accuracy. There was a significant decrease of fair wage ratings from the first to the second half of task completions for the 1-back task ($t = 3.1$, $p = 0.0026$) and 3-detect task ($t = 3.8$, $p < 0.001$). There was no change in fair wage ratings across the first and second halves of experience with the 2-back task ($t = 0.75$, $p = 0.45$). There was no change in accuracy in the first and second halves of task completions on the 1-back task ($t = 0.22$, $p = 0.82$), 3-detect task ($t = -1.8$, $p = 0.076$), or 2-back task ($t = -1.0$, $p = 0.31$). The same was true of mean response times during the tasks (1-back task $t = -0.92$, $p = 0.36$; 3-detect task $t = 0.77$, $p = 0.44$; 2-back task $t = 1.4$, $p = 0.17$). Taken together, these results suggest that any decrease of fair wage ratings over task iterations stems from the experimental design, and not from learning or practice effects. We investigate this further with computational modeling below.

We used the BDM procedure to incentivize subject honesty in their fair wage ratings, so as to accurately measure the costs of cognition. However, one possibility is that, instead of adjusting their fair wage ratings to the demand of each round of each task, subjects instead tried to best the fair wage procedure by matching their ratings to the amount the computer last offered

them. The correlation between subject ratings on trial t and computer offers on trial t-1 was significant in 3 out of 100 subjects at a significance level of $p < 0.05$. This correlation was not significant combining data across all 100 subjects ($\rho = 0.030$, $p = 0.091$). These results suggest that subjects were not engaging in offer-matching behavior.

Analysis of self-report measures

We ran fixed effect regressions on task behavior with linear and quadratic NFC and SAPS terms, using a model selection procedure which trimmed each model down to an intercept term, and the self-report terms which were necessary for model significance ($p < 0.05$). NFC scores were linearly and quadratically related to mean 3-detect accuracy ($\beta = -12$, $\beta = 1.8$). NFC was quadratically related to difficulty ratings for the 1-detect ($\beta = -0.060$). SAPS scores were linearly and quadratically related to mean 1-back accuracy (linear $\beta = 13$, quadratic $\beta = -1.6$), mean 3-detect accuracy (linear $\beta = 7.6$, quadratic $\beta = -0.87$), and difficulty ratings for the 2-back task (linear $\beta = -1.2$, quadratic $\beta = -0.13$). SAPS scores were also quadratically related to 2-back accuracy ($\beta = -1.1$). Neither NFC nor SAPS score were linearly or quadratically related to mean RTs.

We ran the same regression analysis on mean fair wage ratings, collapsed across all tasks. There was a significant quadratic relationship of NFC and mean fair wage ratings ($\beta = -0.028$). We split subjects up into self-report tertiles to further investigate the significant quadratic relationships between task and self-report variables. The tertile split resulted in 25 low, 37 mid, and 37 high NFC subjects, and 34 low, 37 mid, and 28 high SAPS subjects. Post-hoc t-tests confirmed that the significant quadratic effect of NFC is driven by the difference in mean fair wages between the high and mid NFC subjects. Mid NFC subjects had higher fair wage ratings than high NFC subjects ($p = 0.0079$; S3 Fig). However, there were no differences between the low and high NFC groups ($p = 0.15$), or the low and mid NFC groups ($p = 0.30$). We supposed that high NFC subjects would ask for the lowest fair wages, but we did not find such a pattern in explicit fair wage ratings. We next investigated how NFC was related to the implicit costs of cognition captured by our computational model.

Model-based results

Based on the model-agnostic results, we designed and tested a series of computational models to measure the costs of distinct cognitive processes from fair wage ratings. These models allowed us to test the hypotheses that specific cognitive operations are costly to perform, and to estimate the magnitude of these costs. We also measured the costs associated with certain behaviors, including making errors. In doing so, we assessed whether fair wage ratings also captured costs stemming from physical effort (making key presses) or error avoidance, which are not cognitive process costs but are still potential modifiers of fair wages.

We fit subjects' behavior with a series of models using the Computational Behavioral Modeling (CBM) toolbox [32]. All models included a noise parameter (σ), and at least one initial task rating parameter (*init*) as free parameters. One class of models assumed that the cost parameters were fixed across trials, but that the subjects learned about the total cost of completing each task with a learning rate (α). A separate class of models assumed that subjects' demands reflected the cost just on the previous iteration of the task, but that the cost parameters changed linearly with trial number at a rate given by a cost-changing parameter (δ). Within these model classes, we tested several combinations of cost parameters. The maintenance cost ($c_{maintenance}$) captured the effect of maintaining more information in WM. Here, we define the maintenance demand (or WM load) in terms of the number of items subjects are instructed to maintain in WM to complete each task (1 item on 1-back, 2 items on the

2-back, and 2 items on the 3-detect). The interference cost ($c_{interference}$) captured the effect of "lure" trials in the 2-back task. A "lure" trial was defined as one where the stimulus 1 back in WM storage matched the current stimulus on screen, forcing subjects to adjudicate between this false match and true 2-back matches. The update cost (c_{update}) captured the effect of updating WM with new information. Updating WM is unnecessary if the stimuli in memory match the current one on screen, otherwise it happens on every trial. The response cost parameter ($c_{response}$) captured the influence of making button presses, or responding to perceived matches. The miss cost (c_{miss}) captured the effect of making omission errors. The false alarm cost (c_{fa}) captured the effect of making false alarm errors.

We used a model fitting technique which simultaneously leverages similarity between subjects' parameter values and differences in each subjects' best-fitting model to achieve accurate parameter recovery. This technique produces model responsibility scores per subject. Summing these responsibilities across subjects produces model frequency scores, which estimate the prevalence of that model in the overall subject population. The models with non-zero model frequencies in our subject pool included learning rate α, rating noise σ, three initial rating parameters (one per task), and some combination of update costs, interference costs, maintenance costs, false alarm costs, and miss costs (Fig 3A). Two subjects were best fit by a model including the cost-changing parameter δ and a fixed learning rate $\alpha = 1$ but most subjects' (98/100) experienced costs of cognition were stable across 32 task rounds. Most changes in fair wage ratings were likely driven by cost learning (α), differences in the cognitive operations required in different task rounds, or reporting noise (σ).

The model with the highest model frequency included learning rate α and update costs, and was the winning model overall with a protected exceedance probability of >0.99 and a model frequency of 78%. The second most frequent model included interference costs and had a model frequency of 10%. The third most frequent model included update, interference, and maintenance costs, and had a model frequency of 6%. The remaining five recovered models contained the rest of the cost components (including false alarm, miss, and response costs) in various combinations and accounted for the last 6% of model frequency. They also contained two models with δ cost-changing parameters instead of α cost-learning parameters.

Although most of our subjects were best fit by the winning model, one quarter of our subjects were best fit by other models. Subject fair wages were better fit by simulating data for each subject using their best-fitting model (mean $r^2 = 0.52$; Figs 3C and S3), than by simulating data for all subjects with just the winning model (mean $r^2 = 0.47$). In addition, 10 subjects' data were best explained by models containing multiple costs of cognition. Thus, subjects' fair wages were influenced by more than just update costs.

There was scant evidence that button presses or errors were costly, as all models including response, false alarm, or miss cost parameters had a total model frequency less than 3%. Models including response and miss costs each accounted for model frequencies less than 1%, so these costs are not explored further below.

The mean update cost was 0.62 (Fig 3B), making it the highest magnitude cost parameter. The next highest mean parameter value was the interference cost, at 0.60, followed by the maintenance cost at 0.20, and the false alarm cost, at -0.65. Despite the near equivalence of the mean update and interference costs, lures in WM were much less frequent than updates to WM. Because of this, subjects lost more monetary bonuses due to the avoidance of update costs, resulting in their forfeiting an average of 0.87 cents extra per round. They were willing to forfeit 0.26 cents and 0.38 cents per round to avoid maintenance and interference costs, respectively. While subjects did not know the exact mapping between BDM points and the monetary bonus at the conclusion of the experiment (1 point = 1 cent), this speaks to the true costliness of each component process, in terms of the overall monetary amounts subjects forfeited.

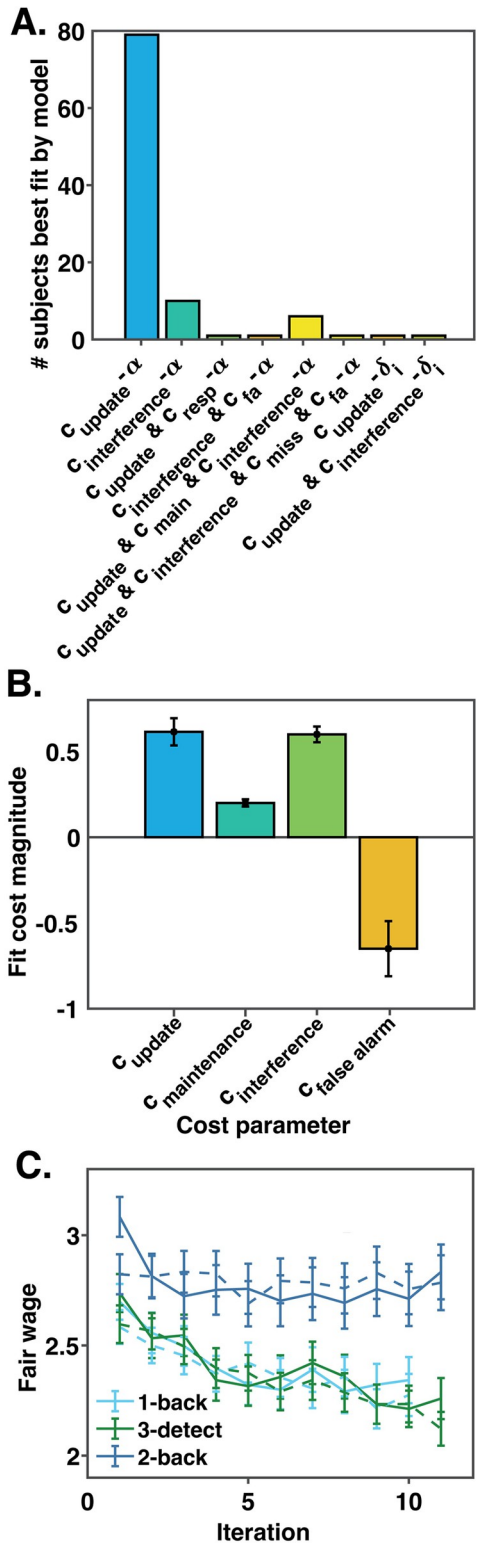

Fig 3. Computational modeling results. A. The number of subjects best fit by each model with a non-zero model frequency. Of the 84 computational models fit to subjects' fair wages, the winning models were alpha cost-learning models containing update costs (c_{update}), interference costs ($c_{interference}$), and maintenance costs ($c_{maintenance}$), and false alarm costs (c_{fa}), in various combinations. The model with the highest model frequency was the model including update costs alone. **B.** The mean of the posterior distribution of each cost parameter from the models that best fit at

least 1 subject's fair wages. These posterior distributions were calculated by combining inferred parameter distributions across subjects and across models. Inference was performed over joint 4D distributions to capture covariance between update, interference, maintenance, and false alarm costs. For plotting purposes we summed over the three irrelevant dimensions for each parameter to construct its marginal distribution, and then calculated the means and variances of the marginals. Error bars reflect the hierarchical standard error of the mean; they were calculated not with the square root of the total number of subjects in the denominator, but with the square root of the number of subjects' data explained by models containing that parameter. Note that the error bars describe the spread of the marginal parameter distributions, not variance in the fitting process, and thus are not suitable for estimating the statistical significance of the effects plotted. **C.** Real (solid lines) versus simulated (dashed lines) fair wage ratings on each rating iteration for each task. Data simulated using each subjects' best model faithfully reproduces real subject data ($r^2 = 0.52$).

https://doi.org/10.1371/journal.pcbi.1012060.g003

As we hypothesized, the mean difference between fair wage ratings on the 2-back and 3-detect tasks was predicted by the magnitude of the interference costs ($r = 0.42$, $p < 0.001$). The mean difference between ratings on the 2-back and 1-back was predicted by the magnitude of the maintenance costs ($r = 0.41$, $p < 0.001$). These correlations confirm that the tasks differ in their subjective costliness at least partially because of the differences in WM operations required by them.

We tested whether any self-report measures of effort avoidance or perfectionism related to fit cost parameters. Specifically, we wondered whether the need for cognition (NFC) or perfectionism (Short Almost Perfect Scale; SAPS) scales were predictive of any cost parameter values. For simplicity, we analyzed just parameter values from subjects best fit by the winning (update costs) model (N = 79). We ran a fixed effect regression including both linear and quadratic terms for the effect of NFC and SAPS scores on fit update cost parameters from the winning model. We found no significant linear or quadratic relationships between update cost and NFC (linear $\beta = 0.22$, quadratic $\beta = -0.044$, full model $p = 0.72$). There was also no significant linear or quadratic relationship between update cost and SAPS score (linear $\beta = -0.21$, quadratic $\beta = 0.027$, full model $p = 0.37$). NFC and SAPS scores were well-sampled across our sample of 100 subjects (S2 Fig).

We then examined whether there were parameter differences across NFC and SAPS tertiles. Within the subjects best fit by the winning model, high NFC subjects had significantly lower update costs than both low ($p = 0.042$) and mid-NFC subjects ($p = 0.040$). There were also differences in initial fair wage ratings across NFC groups (Fig 4), the general pattern being that mid NFC subjects asked for the highest initial fair wages. High NFC subjects had significantly lower initial fair wage ratings than mid NFC subjects for all three tasks (1-back $p = 0.0054$; 2-back $p = 0.0013$; 3-detect $p = 0.022$). There were no significant differences between low and high NFC subjects' initial rating parameters. Mid NFC subjects had higher initial ratings for the 2-back task than low NFC subjects ($p = 0.013$). Mid NFC subjects had higher variance (σ) around their fair wage ratings than high NFC subjects ($p = 0.0071$) and low NFC subjects ($p = 0.014$). There were no significant differences in learning rates between subjects split into NFC tertiles (low vs. high $p = 0.89$; low vs. mid $p = 0.89$; mid vs. high $p = 0.75$). Taken together, these results suggest that both explicit reports about task costliness (initial fair wage ratings for each task), and more implicit experiences of the costs of cognitive operations (update costs) change with individual differences in NFC across subjects.

There were no significant differences in cost parameter magnitudes between subjects split into SAPS tertiles.

Discussion

Deploying working memory or paying attention can feel costly [5,33]. In this work, we quantified the subjective costs of the cognitive operations demanded by commonly studied working

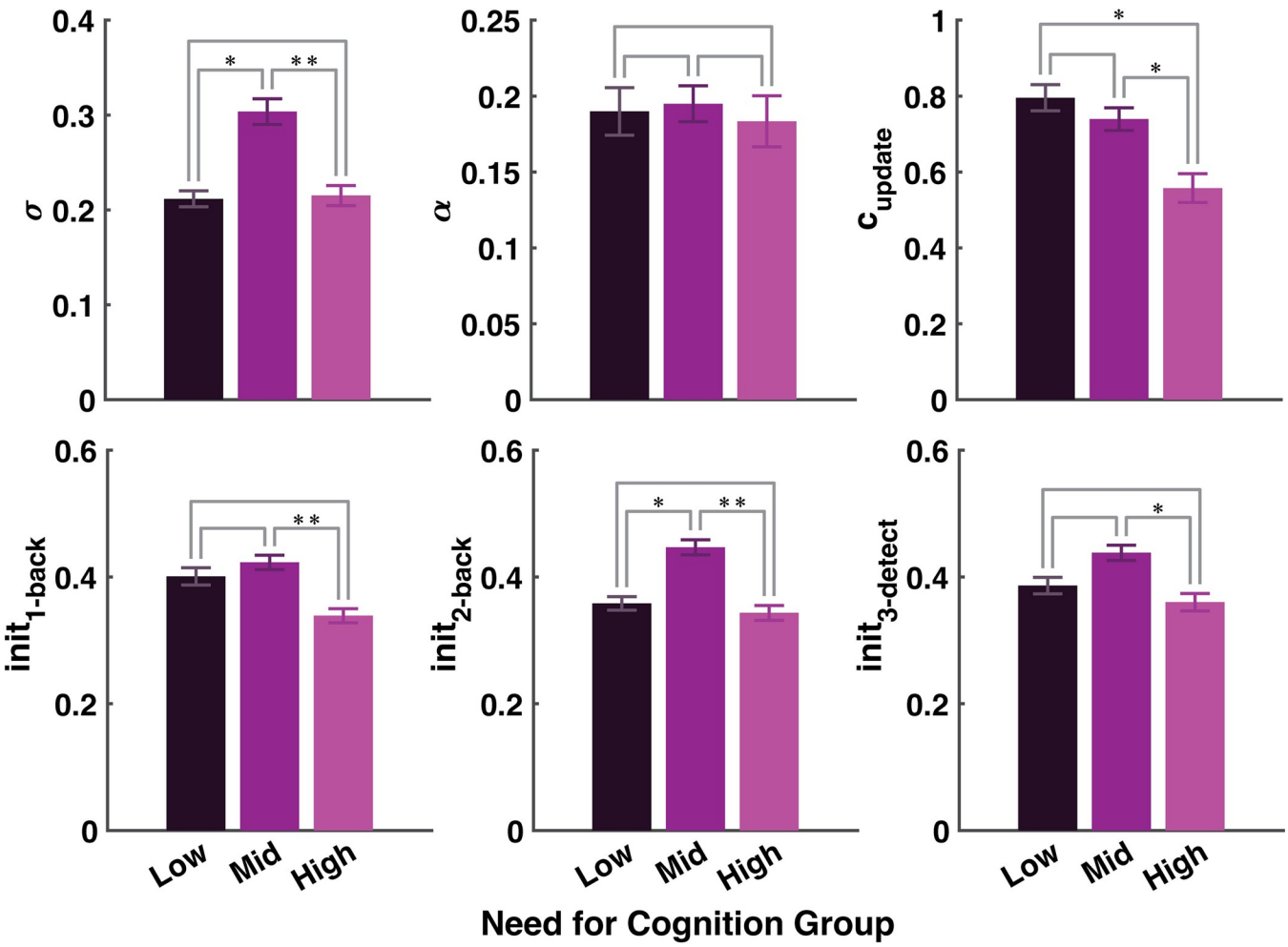

Fig 4. Winning model parameter values by Need for Cognition (NFC) Group. Mean parameter magnitudes from the winning 6-parameter update cost model. σ is the standard deviation parameter which dictates how noisy each subject's fair wage ratings are, on average. α is the subject-specific task cost-learning rate. The update cost is the magnitude of the influence of WM updates on each subject's fair wage ratings. The *init* parameters dictate each subject's initial fair wage for each task. Subjects were split into NFC tertiles resulting in low (N = 25), mid (N = 37), and high (N = 37) NFC groups. Fit parameter values were then averaged within-group to produce each bar. Error bars are standard error of the mean. * indicates significant difference as assessed with a t-test at $p < 0.05$ level. ** $p < 0.01$.

https://doi.org/10.1371/journal.pcbi.1012060.g004

memory and attention tasks, in a way sensitive to both the dynamics of cognitive effort exertion and individual differences in effort avoidance. Using a novel experimental paradigm which leverages an inverted Becker-Degroot-Marschak auction procedure [21], we obtained subject ratings of the total cost of completing a working memory or attention task, one round at a time. We then used a computational model to decompose these ratings into the costs of the individual cognitive operations putatively used during that round, as well as aspects of subject behavior, like errors. Our computational models quantify the subjective costs of individual cognitive operations and allow us to test several hypotheses about how cognitive effort costs may change with time or task experience.

We found evidence that updating WM, interference from within WM storage, and WM maintenance are subjectively costly. Most subjects tracked a single cost. The largest percentage of subjects tracked just update costs, and the next highest proportion tracked just interference costs. Although effortful cognition can be rewarding [3,34], we find that the costs, not the

intrinsic rewards, of cognitive effort drove fair wages. Updating WM cost the most. Subjects forfeit on average 0.87 cents extra per round as a result of avoiding frequent WM updating demands. Interference costs (lure stimuli inside of WM) were similarly high, but because lures were somewhat infrequent, rating them highly (and thereby avoiding them) led subjects to lose less money per round. The third highest cost was that of maintaining more information in WM.

Increasing WM load (the N in N-back) has often been assumed to be the primary driver of increases in subjective difficulty. However, we show that WM load was only minimally costly and that updating and interference had a greater influence on subjective cognitive effort. Other work has shown that WM maintenance demands minimally influence cognitive effort avoidance behavior [17] and do not evoke task performance costs [35]. Lure stimuli in WM storage demand an accurate maintenance of both stimulus identity and stimulus order. The interference cost potentially captures the confusability of stimuli in WM storage and the high subjective cost of disambiguating them by their temporal order. WM updating is similarly complex, as information must be gated in, gated out, and temporally re-ordered. WM updating has been compared to switching between WM attractor states, which could be an energetically costly process [36,37]. Perhaps the magnitude of the update cost parameter captures the complexity of or energetic costs associated with this operation.

We find that subjects quickly learned the costs of completing each task through internal cost feedback signals, then exhibited stable fair wage ratings. Our models provided two surprising new insights into how the costs of cognition may figure into deciding between several paths of action. First, only 10 subjects were best fit by models which contain multiple cost parameters. Tracking multiple costs of cognition may be in itself costly, so subjects may have selected just one cost component to base their fair wage ratings on to minimize overall experimental demands, consciously or otherwise. Second and seemingly at odds with previous work [38–40], we found no evidence that fatigue impacted fair wage ratings as cost parameters did not increase or decrease over rounds. However, cognitive fatigue may only emerge after longer durations of cognitive work [41].

Our task design directly controlled for one possible confound of the costs of cognitive effort, time on task [10–12], by ensuring that the time between trials was the same, no matter the subjects' reaction times. We also standardized the time spent on each round of all tasks. However, subjects' reaction times did vary between tasks, rounds, and trials, and we did test whether the mean reaction time per round was another possible source of effort costs, alongside the other costs of cognition. This reaction time cost was ultimately omitted from our analyses, as it accounted for 0% model frequency in our subject population.

Another key confound in cognitive effort avoidance work is error avoidance [28,29], which is harder to directly control for, as tasks which are cognitively effortful often also elicit more errors. Instead, we measured error rates and their influence over subjects' fair wage ratings. First, there was no relationship between round-by-round accuracy and fair wage ratings in two out of three tasks. Second, while 2/100 subjects' fair wage ratings were responsive to false alarm errors, the fit cost of making false alarms was of the smallest magnitude, and in fact, numerically negative (Fig 3B). Last, only one subject was affected by the cost of making omission errors (misses). These results suggest that error commission is a factor in the overall costs of cognitive effort but certainly is not the only component driving them. However, our analyses cannot control for the generally held belief that the probability of committing errors is high on tasks of great subjective difficulty. That is, perceived error rates, which in our case are related to but not entirely predicted by actual error rates, may have influenced subjects' fair wage ratings in a way we could not measure. In addition, just because error commission does not influence fair wages, does not mean that error avoidance does not influence them.

Cognitive operations, like protecting against WM interference, may evoke subjective feelings of effort partially because subjects are estimating the potential for making an error if they fail to properly apply them. Therefore we cannot completely rule out the involvement of error avoidance in the reporting of cognitive effort costs.

The Need for Cognition (NFC) scale measures the self-reported tendency to engage in challenging cognitive work [2]. Our task and modeling approach may be sensitive to self-report NFC, as the cost of updating WM is lowest in subjects with high NFC. This suggests that what we measure with our paradigm is related to the trait tendency to avoid cognitive effort [42]. However, our study was not designed nor sufficiently powered to adequately establish the ability of our paradigm to capture self-reported individual differences. For one, the same NFC scores which linearly related to the cost of updating WM exhibited a quadratic relationship with initial task ratings. Second, though one would suspect that high NFC subjects would on average provide the lowest fair wage ratings, their fair wage ratings were not significantly different from low NFC subjects' ratings. Lastly, there were no relationships between perfectionism (SAPS) scores and task- and model-based measurements. Therefore, whether our paradigm is sensitive to individual differences should be considered an avenue for further investigation.

One limitation of our task design was the high degree of correlation between cost components, which may have impacted cost parameter recovery during model fitting. While maintenance demands were constant across the 2-back and 3-detect tasks, the 2-back was the only task which required subjects to filter out interference from lures stored in WM. In addition, as the 2-back was the most difficult task, it was associated with the most errors. Thus the total cost components increased from the 1-back to the 2-back, and to some extent from the 3-detect to the 2-back. This resulted in high correlations between cost components within subjects. Despite this consequence of the experimental design, there remained a high degree of fidelity in parameter recovery (S4 Fig), and a low degree of tradeoff between fit parameter values (S1 Text). It remains an open question as to what extent these cognitive operations (i.e. WM updating, resistance to interference, and maintenance) depend on overlapping or independent mechanisms, and indeed whether the costs of these operations are related.

This work directly quantifies the costs associated with the cognitive operations required in working memory and attention tasks, not just how subjects avoid or approach each task. The N-back, a classic WM task, is useful in the study of working memory because it requires the use of many diverse WM operations [31]. Here, we reveal that the N-back's strength may also be its weakness, in that the number of WM operations required to complete it is also what makes it so aversive [43].

There are many avenues for future work using this experimental and modeling approach. For example, the original Demand Selection Task (DST; [4]) measured demand avoidance in response to varying frequencies of task-switching. Kool et al. reported a demand learning curve of a very similar shape to that which we report here. Our fair wage rating and modeling procedure could be used to quantify the subjective costs of task switching, in terms of monetary value. However, as the original DST is not explicitly incentivized, it is worth testing in future work whether a non-incentivized rating procedure (i.e. plain "effort" or "difficulty" ratings) would work similarly well.

Here, we adopted one specific process model to decompose each round of each task into the component cognitive operations necessary to complete it, though there are many possible models to use. The use of a different process model could have resulted in a different cost component structure. Additionally, it is possible that some subjects used a different strategy than simple WM storage on any of our tasks. While outside the scope of this work, future work using our approach could also ask subjects to report whether they used any cognitive load-

reducing strategies during the tasks. These reports could then be used to root the process model in subjects' self-reported strategies for task completion.

These results have potential implications for treating cognitive dysfunction in psychiatric disorders. For one, the N-back task may not be suitable for use as a benchmark for WM ability in psychiatric populations, as many have comorbid cognitive and motivational deficits. Dopaminergic cortico-striatal loops, which are highly sensitive to reward, are thought to be a driver of WM performance [44–46]. Our novel paradigm may be clinically useful, as cognitive dysfunction could be partially treated by comparing the costs of cognition across groups, then offsetting those costs with rewards [47–49].

In summary, along with a novel experimental approach in which subjects request wages for completing one round of one task, we implemented a modeling procedure that decomposes their wages into the costs driving them. We found that updating WM, interference among items in WM, and WM load are costly, above simple error, time, or fatigue costs. This suggests that certain cognitive operations are inherently costly to perform, in alignment with the idea that human cognition is subject to cost-benefit analyses which can result in the use of less costly, less effective cognitive strategies [50]. Surprisingly, the highest subjective cost of N-back performance was not WM load, but WM updating. We also find a small, but significant relationship between self-report individual differences in cognitive effort avoidance and the implicit costs associated with WM updating. Often, task-based and trait measures of effort avoidance do not relate [7,42,51,52]. While further work and a larger sample are needed to confirm the strength of this relationship, this preliminary evidence suggests that our model parameters may relate to dispositional individual differences. Our task and modeling paradigm may therefore be useful in psychiatric or developmental populations to measure, then offset, the costs of cognition.

Methods

Ethics statement

Participants provided written informed consent in accordance with procedures approved by the Ethics Committee of the Medical Faculty and Medical Clinic at the Eberhard-Karls-University of Tübingen (approval number 734/2019BO1).

Subject sample

100 subjects (35 female, 14 unspecified sex, mean(std) age: 39 [12], 11 unspecified age) completed our online task in full. 281 unique workers opened our experiment on Amazon Mechanical Turk (AMT). Of the 270 subjects who consented to participate, 218 of them made it through the practice blocks, 142 successfully finished the quiz, 125 made it to the 16th block of the experiment, and then 100 completed the experiment in its entirety. Our final sample, which we analyze below, consisted of these 100 subjects who finished the experiment. We did not include any data from any of the subjects who did not finish the experiment in our analyses. Given the strict accuracy and attention cutoffs we imposed, and the overall length of our task (mean(median) total time on task: 37 [36] minutes) versus the typical length of tasks on AMT (one study reported that the mean time spent on submitted HITs was less than 2 minutes [53]), we considered a 37% completion rate to be acceptable.

Experimental procedure

Subjects were asked to complete 32 task rounds, alternating between 4 different tasks: a 1-detect task (oddball detection), a 1-back task, a 3-detect task (detect 3 of the same stimulus

in order), and a 2-back task. We chose these four tasks because they rely on many of the same cognitive processes, whilst also differing in important ways in the operations they require from those processes. All four tasks require subjects to attend to the screen while letters are presented one at a time in order to search for specific targets. Subjects were tasked with making a button press every time they saw one of these targets. The targets changed for each task, thereby changing the cognitive demands of each task. In the 1-detect task, the target was the letter "T". In the 3-detect task, the target was any letter presented 3 trials in a row. In the 1- and 2-back tasks, target trials were ones in which the letter on screen matched the letter displayed 1 or 2 trials back, respectively.

In a novel experimental paradigm, we leveraged the Becker-Degroot-Marschak (BDM) auction procedure to measure the evolving subjective value of choice options [21]. The experiment was coded using a pre-built Javascript framework for online Psychology experiments (JsPsych; [54]) and custom Javascript functions. Subjects were introduced to 4 tasks, each of which was associated with a fractal image (a "task label"; see Fig 1): the 1- and 2-back working memory tasks, and two types of attentional vigilance task, which we refer to as the 1-detect (the default task) and 3-detect [4,5,19,31]. The task label image was presented during the initial task instructions along with the following text: "This picture will always be associated with the following task, like a picture label." To ensure that subjects learned to associate each task label with its paired task, the fractal remained in the upper right corner of each trial of the task.

Following practice, (i.e. in the main experiment) subjects completed a total of 32 rounds, using the BDM procedure before each round to report the wages they considered fair for performing the particular non-default task that was offered instead of the default 1-detect task. In all tasks, subjects saw a sequence of 15 letters, one after the other. Subjects had to respond to the letters that matched a rule by pressing the "K" key on the keyboard. Stimuli remained on the screen for 1.5s; any response had to be made before they disappeared. If a subject responded late to a match, that trial was marked incorrect. The inter-stimulus interval was 300ms. Time on task was standardized such that the time spent on each task could not influence subjective effort cost differences across tasks; each task round took approximately 24 seconds.

The 1-detect task was the default task, intended to involve minimal effort. Subjects had to respond only if they saw a "T" on screen. In the 3-detect task, subjects had to respond when any letter was presented 3 trials in a row. In the 1- and 2-back tasks, subjects had to respond when the letter on screen matched the one displayed 1 or 2 trials back, respectively. Letter sequences were standardized such that subjects were required to respond to 3 to 5 matches per round, regardless of task identity. Therefore targets were presented on 20% to 33% of trials. Other non-target letters were presented on the remainder of trials. All stimuli were presented one at a time. We chose to run these four tasks because they involved similar cognitive processes, but differed in their rule structure and thus the number and complexity of the operations they required. In particular, we sought to measure the costs of increased WM load and the information manipulation required by the N-back tasks.

Comparing the subjects' fair wage demands for the 1- and 2-back tasks allowed us to measure the cost of maintaining one more item in working memory ("maintenance"). Comparing the demands for the 2-back and 3-detect tasks, which both require the maintenance of 2 items, allowed us to measure the cost of protecting against interference in the contents of WM ("interference"). In the 3-detect task, subjects had to remember the 2 previous stimuli and compare them to the current stimulus. Detecting a match was simple as long as one recalled whether the previous 2 stimuli matched the current one. In the 2-back task it remained essential to recall the previous 2 stimuli. However, the stimulus from 1 trial ago was never relevant for the trial at hand; all that mattered was the identity of the stimulus 2 trials ago. Because both

must be stored, however, it is possible that the stimulus from 1 trial ago was distracting, and if it matched the current stimulus, it may have served as a lure to respond. Identifying and filtering out this distraction may require significant attention and effort. Thus a "lure trial" was any trial where the irrelevant stimulus from 1 trial ago matched the stimulus on the current trial, in the 2-back task. The interference cost in our model captured the cost of these lure trials. This idea is also described in [55].

Stimuli were presented in pseudo-random order such that the use of other WM operations, like WM updating, also differed slightly across rounds. Additionally, forcing 3–5 matches per round allowed us to measure the costs of responding to perceived matches, not responding when matches occur (misses), or responding erroneously (false alarms).

To obtain fair wage ratings for each round in each task, we employed an inversion of the typical Becker-Degroot-Marschak (BDM) auction procedure, in which subjects bid for items with points. In our procedure, subjects are asked to do some cognitive work in exchange for a fair wage. Before each round, subjects were shown a fractal image associated with one of the tasks, and were asked to use a slider to specify their "fair wage" for completing one round of that task. Possible fair wages ranged from 1 to 5 points. They were then shown a random computer offer, also from 1 to 5 points. If the computer offer was above their requested wage, they were given the computer offer for completing one round of the task associated with the fractal. If the computer offer was below their requested wage, they completed the default (1-detect) task for 1 point. All task rounds consisted of 15 trials.

We used the BDM procedure in this work because, via mechanism design, it motivated subjects to report the true subjective value of the effort they expected to expend on each instance of a task. If subjects were effort avoidant and wanted to earn higher wages or not complete effortful tasks at all, they would ask for high wages. If subjects were effort seeking, or at least not effort avoidant, then their fair wages would be low as they should be satisfied with any number of points above the minimum. If one task was substantially more effortful, then our subjects should ask for higher wages on that task so that they would not have to complete that task without proper compensation. In an attempt to prevent subjects from being overly avoidant of making errors, we did not impose an accuracy cutoff for the receipt of points on individual rounds. Further, after the initial practice phase, subjects were not informed of their accuracy each round. However, subjects were aware that if they were inattentive to the task, or their overall accuracy fell below some cutoff, that the task would conclude early and they would receive less compensation (see exclusion criteria below). At the end of the task, subjects' points were tallied and converted into a monetary bonus.

At the end of the main experiment, subjects completed a basic demographic inventory, the Need For Cognition Scale (NFC; [2]), and the Short Almost Perfect Scale (SAPS; [30]). They then rated the difficulty of each of the tasks (signaled by its associated fractal) using the same slider that they used to provide their fair wage ratings. Subjects were also able to provide comments on their experiences completing the experiment. Subjects were given one hour and 15 minutes to complete the entire experiment.

Recruitment and exclusion criteria

The subject pool was limited to Amazon Mechanical Turk workers based in the United States, to ensure English reading comprehension. We limited our recruitment to workers ages 18 and up with at least 100 completed Human Intelligence Tasks, and with at least 85% acceptance rates. We also ensured that subjects had not completed the task before using their Worker ID. To ensure that subjects understood the task and were able to maintain a high level of accuracy, we excluded subjects who did not demonstrate task proficiency or an understanding of the fair

wage procedure after the practice phase. We implemented two tests that subjects had to pass to make it into the main experiment. First, subjects had to reach 80% task accuracy on 15 trials of our most difficult task, the 2-back. They had up to 10 rounds to do so. 52 subjects failed to reach this criterion. Following that, subjects had to correctly answer 4 out of 6 questions about the BDM procedure. 76 subjects did not pass this quiz. If subjects passed both those checks, then they proceeded to the main experiment. After these exclusions, 142 subjects started the main experiment.

During the main experiment, subjects' performance was assessed 3 times (every 8 rounds). If in 8 rounds, subjects missed the response deadline for 4 fair wage ratings or their overall accuracy went below 60%, the task ended early and their data were not used in the final analyses. This eliminated another 42 subjects, resulting in a sample size of 100 subjects total. Subjects were given a 30 second rest between task rounds and no other breaks.

Model-agnostic analyses

All model-agnostic and model-based analyses were run in MATLAB [56]. Subject accuracy was calculated online as a weighted function of correct responses (hits) and correct withholding of responses (correct rejections), where hits were given three times more weight than correct rejections. We chose to emphasize hits over correct rejections in order to encourage participant engagement in the tasks, though subjects were not aware of the exact scoring procedure. In this way, subject accuracy was tracked while they completed the experiment, so that subjects who were not engaging with the task could be removed from the experiment early. Once subjects completed the experiment, we examined their behavior on each task by running ANOVAs on accuracy, response time, fair wages, and difficulty ratings, looking for an effect of task identity. We examined significant main effects of task identity with post-hoc t-tests. We correlated mean accuracy on each task and mean fair wage for that task across subjects. Additionally, we ran linear fixed effect regressions on accuracy versus task iteration and overall experimental round, to examine potential learning or fatigue effects on accuracy. We ran these same analyses on fair wage demands to determine whether subjects' fair wages changed with time or task practice. Additionally, we ran a comparison of reaction times on fair wage ratings at the start and end of the experiment.

To investigate possible offer-matching behavior, we ran two analyses relating subjects' fair wage ratings to the randomly generated offer previously presented to them. First, we correlated each subjects' fair wage rating on round t with the computer's randomly generated offer on round t—1. This may not be the most robust analysis within-subject due to the low number of ratings per subject overall [32]. To increase our sensitivity to any possible effects, we also correlated fair wage ratings and the previous computer-generated offers across all 100 subjects. In both analyses, we used a non-parametric Spearman correlation as neither variable was normally distributed.

We scored subjects Short Almost Perfect Scale (SAPS) and Need for Cognition Scale (NFC) responses by summing the numerical values of all their answers, reversing some values as indicated by published scoring guidelines, then dividing by the number of questions answered. We used this normalization to ensure that any questions that were not responded to would not artificially lower questionnaire scores. We excluded questionnaire data from subjects who incorrectly answered one or both of our screener questions (i.e. "Please select 'Strongly Agree' for this question"). This type of attention check has been shown to be a reliable way of removing subjects who are randomly responding to questionnaires, especially when administered more than once during an experiment [57].

The mean(std) normalized NFC score was 3.4(0.9) and the mean(std) normalized SAPS score was 4.4(1.3). 1 subject chose not to finish those questionnaires and as such has no NFC or SAPS score.

We correlated these questionnaire scores with each other and with participant age. NFC and SAPS scores were positively correlated ($r = 0.24$; $p < 0.05$). There was no relationship between participant age and NFC (NFC/age $r = -0.17$; $p > 0.1$) or SAPS score (SAPS/age $r = -0.16$; $p > 0.1$). We also regressed NFC and SAPS scores, and their squares, against mean fair wage ratings, average accuracy, and average response time. We used a model selection procedure which reduced each fixed effect regression to an intercept term, and the self-report terms which were necessary for model significance ($p < 0.05$). If two reduced models were significant, and they included different terms, we selected the model with the lower mean squared error (MSE) in predicting each task variable. We did this to assess both linear and quadratic relationships between individual difference scores and task performance measures.

To build upon the quadratic relationships observed, we also split subjects into tertiles based on their questionnaire scores. Because both scales administered were short-form, many subjects have the same score. Thus after splitting subjects into low, mid-, or high scoring groups based on these scores, the resulting tertiles did not have the same number of subjects in them. Nevertheless, we ran a series of ANOVAs and post-hoc t-tests to examine whether these groups differed in their task accuracy, or fair wages.

Computational methods and model-based analyses

We used a computational model to quantify the putative cognitive processes used in task completion and their influence on fair wage ratings. We used a process model to decompose each task into the cognitive operations putatively involved in its completion. Each model included one σ parameter to describe rating noise, at least one initial task rating parameter (*init*), one α or δ parameter, and at least one cost parameter. We made no assumptions about which combination of cost parameters would best fit subject data, and so tested models including all possible combinations of cost parameters. However, not every model we specified was included in the final fit, as we limited model fitting to those models with high individual parameter recoverability. We fit 84 candidate models to subject data. This number is elevated by our use of two different functions of how fair wages change with time.

We modeled subjects' fair wage ratings as a dynamic process driven by subject learning (with learning rate α) or by the changing costs of cognitive effort (with cost changing parameter δ_j). The first class of models tests the hypothesis that the total cost associated with each task is learnt through experience with the task and the number of costly components required to complete it. The second class tests the hypothesis that cognitive effort costs may themselves change over time, as costly processes become either less costly with practice or more costly as subjects grow fatigued. We also tested a third class of models combining these mechanisms (S2 Text).

The fair wage ratings for each task were initialized in the model by fitting initial rating parameters for each task and each subject, thus capturing each subjects' initial ratings with very high fidelity (S4 Fig). Each subject's initial fair wage ratings for each task were captured using a free parameter $init_i$.

$\text{rating}^0(\text{task} = \text{"1-back"}) = init_{1\text{-}back}$;

$\text{rating}^0(\text{task} = \text{"2-back"}) = init_{2\text{-}back}$;

$\text{rating}^0(\text{task} = \text{"3-detect"}) = init_{3\text{-}detect}$

We presume that these initial fair wage ratings reflected subjects' subjective experiences of the costs of each task during the practice phase. These estimates were likely noisy because

subjects had very little task experience before making their first fair wage ratings. Due to a data saving error in our first few subjects, we did not relate these initial ratings to the task features (putative costs) present in the practice phase. We also cannot speak to the demand for specific cognitive processes during the practice round and whether subjects experienced each possible cost of cognition during practice.

While the inclusion of three extra free parameters to determine initial fair wages may seem over-specific, correctly capturing each subject's starting point allows us to fit most accurately how subjects' fair wage ratings evolve over the course of the experiment, as well as how they respond to individual cost components. However, because there are already extra parameters in the δ_j class of models (+1 δ_j for each cost parameter, so that they can change independently), we did not fit individual *init* parameters to each task in this class of model, to avoid overfitting. After the initial fair wage ratings, the total cost on task round r_k of task k was then used to determine the fair wage rating on the next round of that same task (round $r_k + 1$). This round may arise some trials later; we denote trials by t.

We approached cost decomposition with a simple program which was capable of accurately completing each task with the same "cognitive" functions, but switched between rule structures depending on the task at hand. This program centered on a process model of working memory which was inspired by other approaches to cost decomposition on working memory and cognitive control tasks [35,55,58]. Like these other approaches, ours was theoretically based in a neurally-inspired network model of WM in which WM toggles between flexible updating and stable maintenance modes via interactions between the prefrontal cortex and basal ganglia [36,46,59–61]. In our process model, WM gated in new stimuli and gated out old stimuli simultaneously, a process we call "updating." It held as many items in memory as were necessary to complete each task accurately: we define the WM load to be 0 items on the 1-detect task, 1 item on the 1-back task, and 2 items on the 2-back and 3-detect tasks. We did not test any process models where, instead, a counting strategy was used on the 3-detect task, theoretically resulting in a lower WM load. Instead, we made the simpler assumption that the same basic strategy—update WM, maintain items in order, then update again—was used on all tasks. We also used this process model to keep track of the properties of the stimuli maintained in working memory, including whether they might invoke interference due to the nature of the task. One type of stimulus which was of interest to us was a 2-back "lure" stimulus, or a stimulus 1 trial back in memory which matched the stimulus currently on screen. We hypothesized that this type of stimulus might evoke interference in WM, and therefore increase the costs of cognition by increasing the subjects' need to adjudicate between true 2-back matches and these "lures."

We tallied each operation that this process model had to use to complete each task round with 100% accuracy, including how many items had to be maintained in WM, how many times WM storage had to be updated with new information, or how many times there were interfering "lure" stimuli in WM storage. Cost decomposition was achieved according to a perfect WM process model, so all items presented were considered to be maintained in WM, with no forgetting or noise. Items in WM storage could therefore be marked as both a maintained item and a lure item (interference item), if appropriate. Mistakes were not a part of our process model. Instead, we tallied the mistakes (misses and false alarms) and button press responses made by each subject in each round, and used these as additional factors. It is important to note that a different choice of process model could result in a different cost structure, and we did indeed test a few different process models. To save on computational time and complexity, we included only the costs defined by the process model above in our final model fit. We did, however, run testing on some of the assumptions of our process model, to confirm the superiority of this model relative to other alternatives (see S2 Text).

To model subjects' fair wage ratings, these cost components were scaled by their associated costs (which might change over trials t, and were fit through the modeling), then summed to produce the total cost incurred on that round of that task. For round $r = r_k$ of task k:

$$cost^r(k) = \sum_{j \in C_{params}} components_j^r c_j^t \quad (1)$$

The most complex model included six cost parameters (set C_{params}): the cost of responding to a perceived match ($c_{response}$), the cost of maintaining information in WM ($c_{maintenance}$), the cost of protecting against interference in the contents of WM ($c_{interference}$), the cost of updating WM with new information (c_{update}), the cost of false alarm responding when there was no match (c_{fa}), and the cost of missing a match (c_{miss}). Other than the interference cost, which was only present in the 2-back task, each cost was fit from ratings of all 3 rated tasks. We also tested a version of interference costs which allowed for interference to arise in the 3-detect task, but found that it provided a worse, complexity-controlled fit to subject data (S2 Text). Therefore we opted for the simpler, 2-back-only formulation of interference costs. Again, the other cost components were assumed to account for ratings on all 3 tasks, though of course to differing degrees across tasks. For example, false alarm errors were committed during all three rated tasks, but to the greatest extent during the 2-back task, due to its overall difficulty. All cost parameters were unbounded such that they could be positive, or negative. If any components were perceived to be rewarding, instead of costly, then our model would capture that with a negative cost magnitude.

We tested two possible fair wage rating updating mechanisms: a class of model which assumes subjects learn the stable costs of completing each task through task experience, and a class which allows the costs subjects are learning to be dynamic (i.e. changing due to fatigue). These updating mechanisms are subtly different, and involve two different free parameters: δ, the scalar with which costs are changed trial-by-trial, and α, the cost learning rate. It should be noted, however, that it is theoretically possible that both mechanisms contribute to cost ratings simultaneously. For simplicity and for robustness of model recovery, we chose to fit these updating mechanisms as separate model classes. However, in a supplementary analysis, we also fit these two updating mechanisms jointly, to compare this joint α-δ class of models to the simpler α-only models (S2 Text).

In the α-only version of the model, the costs do not change with trial number, as they do in the other class of models, so: $c_j^1 = c_j^2 = \ldots = c_j^T$. This class of models learns incrementally and α is the subject-specific cost learning rate which captures how much each subject adjusts their ratings for an individual task k based on the most recent round $r = r_k$ of that task:

$$\bar{rating}^{r+1}(k) = \bar{rating}^r(k) + (cost^r(k) - \bar{rating}^r(k)) \quad (2)$$

We modeled noise in the fair wage rating process with a Gaussian noise process centered on 0 with standard deviation σ, also a free parameter, and by applying this noise to each fair wage rating independently. This makes the generated rating follow:

$$rating^{r+1}(k) = \bar{rating}^{r+1}(k) + N(0, \sigma^2) \quad (3)$$

In the cost-changing class of models, δ_j ($j \in C_{params}$) is the cost-specific change parameter which captures how costs linearly change over time (trial number t), i.e. with task experience

or fatigue:

$$c_j^t = c_j^0 * \left(1 + \frac{\delta_j * t}{T}\right)) \quad (4)$$

T is the total number of rounds across the entire experiment, and δ can be positive or negative. The flexibility of δ allows the cost of each cognitive operation c_j to increase or decrease linearly. Note that because the cost components are shared over tasks, and fatigue is supposed to generally increase with time on task, in this model class each cost is changed according to overall trial number (t), instead of task round number (r_k for task k). In this class of models, fair wage ratings on round $r_k + 1$ of task k are a direct function of the cost parameters and task components involved to complete the previous task round $r = r_k$ (which is equivalent to having a cost learning rate $\alpha = 1$):

$$\bar{rating}^{r+1}(k) = \bar{rating}^{r}(k) + (cost^{r}(k) - \bar{rating}^{r}(k)) \quad (5)$$

In the joint -δ class of models, the cost update was scaled by cost learning rate parameter α:

$$\bar{rating}^{r+1}(k) = \bar{rating}^{r}(k) + (cost^{r}(k) - \bar{rating}^{r}(k)) \quad (6)$$

The joint α-δ class of models included both a linear (as described by Eq 4) and an exponential cost-changing function of trial number. The exponential cost-changing parameter δ described the shape and magnitude of the change with trial number:

$$c_j^t = c_j^0 * t^{\delta} \quad (7)$$

Again, we allowed δ to vary between negative infinity and positive infinity, allowing for total flexibility in the shape and magnitude of the cost-changing function as the experiment progressed. The shape of the fair wage rating curves most closely resembled square-root functions, and we therefore expected δ to fall between -1 and 1.

Note that we included these α-δ models in a supplementary model fitting procedure to assess their fit against the α-only class of models, and as a possible replacement to the linear δ-only models. In this supplementary analysis, we tested 112 total models, including the α-only models, which we also fit in the analyses we describe in the Results section. However, the α-δ class of models did not survive complexity-controlled model comparison, and so we describe the results of this exercise only in S2 Text.

Given the modest number of ratings provided by each subject (32 in total, split amongst 3 tasks), and the overall similarity of ratings between subjects, we fit our models using a hierarchical Bayesian inference (HBI) for computational behavioral modeling (CBM) package [32]. Employing a hierarchical parameter estimation procedure allows for similarity across subjects to be leveraged to fit individual parameter values accurately, especially when fitting few individual data points. The package leverages estimations of group parameter means and variances in the individual parameter estimation process. In addition, this package allows for the possibility that not every subject is best fit by one model. Model responsibilities were calculated subject-by-subject such that subjects who were not well-described by a model did not influence the overall parameter probability distributions from that model. In our case, this allowed for individual differences in what processes were perceived as costly. If the ratings of some subjects were not affected by a certain cost term, then the group-level estimate of this cost was not driven down by their inclusion in the pool.

Model responsibility scores are an estimate of the probability that the data for a given subject n were generated by a process described by model k, obtained iteratively by the

hierarchical Bayesian fitting procedure [32,62]. One measure of model fit which we used for model selection, model frequency, is the sum of these subject-by-subject model responsibilities. Model frequency is the sum of model responsibility scores (which range from 0 to 1 within-subject) across all subjects. Model frequency scores can range from 0 to N (N being the number of subjects in the sample). For example, if 3 subjects are not best fit by model X, but the fractional model responsibility for each subject is 35%, then model X has a model frequency of 1.05. In further analyses, we focused only on models with model frequency of at least 1 (in our example, model X would be included in these further analyses). The other measure of model fit we used was the protected exceedance probability [62], which is the probability that model k best describes the data at the group-level, over and above a null (random) model.

The Bayesian model fitting procedure constrains the group parameters to have Gaussian distributions, and so, as is common, we transformed the parameter associated with the learning rate α using a logistic sigmoid (so it lies between 0 and 1) and the parameter associated with the rating noise σ using an exponential (so that it is positive with a log normal distribution).

To assess the winning models' ability to reproduce the behavioral effects of interest, we simulated fair wage ratings using each of the winning models. We then compared these model simulations to real subject behavior via visual inspection, and by computing mean r-squared values for each model. Because stochasticity is one feature of model behavior (via the standard deviation parameter σ), we simulated each subject's data using their fit parameter values 10 different times to control for the stochasticity of these simulations. Each time, we correlated the true fair wage ratings of all subjects with the set of simulated fair wages, and then squared the r-value obtained. We ran this over 1000 iterations, and then took the average to produce a mean r-squared value for each model. This was then used to validate that the models could reproduce subject behavior.

In the CBM toolbox, the group-level mean for each parameter is calculated separately for each model. This allows group-level cost parameter magnitudes to be compared within-model, but not across-model. In order to compare the magnitudes of the cost parameters across all our models, we constructed posterior probability distributions over the magnitude of each cost. We used parameter estimates from every subject and every model, weighing the contribution of each subject s and model m by their fit responsibility ρ:

$$P(\theta|D^s) = \sum\nolimits_m \rho^s_m P(\theta^s_m|D^s)P(\theta_{\tilde{m}}) \tag{8}$$

where $P(\theta_{\tilde{m}})$ is the group-level prior distribution over the cost terms in other models, but which are left out of model m. This prior is a weighted average over the group-level parameter distributions derived from each model, where the weights are again derived from the model responsibilities ρ^s_m. We assumed that these prior distributions were Gaussian within-model, then averaged them across models to produce non-Gaussian mixture models of across-model priors.

Using Eq 6, we constructed a 4D distribution over the four cost parameters included in models with model frequency scores of at least 1. We summed over the 4D joint distribution to produce the marginal distributions of each cost. Additionally, we subdivided our subjects into tertiles based on self-report scores (NFC and SAPS), and calculated the posterior distribution over cost parameter values associated with each score group g:

$$P(\theta; g) \propto \prod\nolimits_{s=1}^{S_g} \left(\sum\nolimits_m \rho^s_m P(\theta^s_m|D^s)P(\theta_{\tilde{m}})\right) \tag{9}$$

where subjects 1 through S_g belong to the group of interest. We then assessed the overlap across groups by comparing the group posterior distributions.

We obtained the means and standard deviations of the marginal posterior distributions over individual cost magnitudes. In this way, we assessed the degree to which the cost magnitudes were separable within- and across-subjects, and across models which did not share all the same parameters.

We confirmed the validity of our models and model fitting procedure by running a generate and recover procedure. For each model, we simulated a data set of artificial subjects with known parameter values. We used trial-by-trial cost components taken directly from subject behavior to ensure that real responses, including errors, and task characteristics were compatible with our modeling procedure. As the fitting procedure first proceeds subject-by-subject, we sometimes ran this generate and recover procedure using 30 or 50 simulated subjects instead of the full 100 subject dataset to reduce fitting time. In these cases, to ensure similarly adequate recoverability across models, we included the same subjects in the smaller dataset for each model such that subject-specific deviations in goodness-of-fit did not unduly influence our assessment of the recoverability of one model in comparison with another. To determine which models were sufficiently robust in parameter recovery, we ran this generate and recover for all 126 possible models (combining different costs and using an α or δ update mechanism). In this way, we selected 84 models to test that showed reliable parameter recovery and minimal cost parameter tradeoff. We wanted to test a broad array of models since we had limited a priori knowledge of which cost components would drive fair wages, or what form cost updates would take. At the same time, we wanted to fit real subjects' data only with models that had recoverable free parameters and minimal tradeoff between costs, despite possible correlations of cost components, as individual differences were of particular interest.

S4 Fig shows the results of this generate and recover procedure for one example model, which includes update, maintenance, and false alarm costs (N = 50 simulated subjects). All fit and real parameters were highly correlated (σ r = 0.84, $p < 0.001$; α r = 0.94, $p < 0.001$; update costs r = 0.55, $p < 0.001$; maintenance costs r = 0.66, $p < 0.001$; false alarm costs r = 0.78, $p < 0.001$; $init_{1\text{-back}}$ r = 0.88, $p < 0.001$; $init_{2\text{-back}}$ r = 0.96, $p < 0.001$; $init_{3\text{-detect}}$ r = 0.92, $p < 0.001$). This indicates that our models supported the reliable recovery of individual parameters, despite the modest number of trials that were fit per subject.

Supporting information

S1 Fig. Mean fair wage ratings across task rating iteration. All subjects completed 10–11 ratings of each task, but only between 1 and 11 rounds per task. Here we plot the mean fair wage for the subjects who completed 1 to 11 iterations of each task, grouped by the total number of iterations they completed. Subjects who completed more task iterations are plotted in darker colors. This illustrates the diversity in fair wage ratings for each task across subjects, as well as the stability of the ratings subjects gave to each task. In addition, it shows that, due to the design of our task, subjects who asked for high fair wages on one of the tasks did indeed complete fewer iterations of that task. Error bars are drawn with standard error of the mean.
(TIF)

S2 Fig. Self-report scores and their relationships to mean fair wages. A. Distribution of Need for Cognition (NFC) scores within the experimental sample. Scores have been normalized by the number of questions answered such as not to lower the mean of the distribution artificially. The distribution of NFC scores in our sample is right-skewed compared to the typical distribution of NFC scores. **B.** Distribution of Short Almost Perfect Scale (SAPS) scores. Scores have been normalized by the number of questions answered such as not to artificially

lower the mean of the distribution. The distribution of SAPS scores in our sample is typical of both in-person samples and other samples on MTurk. **C**. NFC scores versus SAPS scores. NFC and SAPS scores were positively correlated ($r = 0.24$; $p < 0.01$). **D.** Mean fair wage rating on the 1-back, 3-detect, and 2-back tasks by tertile split NFC groups. Error bars were drawn using the standard error of the mean (SEM). There was a significant quadratic relationship of NFC and mean fair wage ratings ($\beta = -0.03$). Post-hoc t-tests confirmed that the significant quadratic effect of NFC was only driven by mid NFC subjects having significantly higher fair wage ratings than high NFC subjects ($p < 0.01$). **E.** Mean fair wage rating on the 1-back, 3-detect, and 2-back tasks by tertile split SAPS groups. Error bars were drawn using the SEM. A 3-way ANOVA, revealed no effect of SAPS group ($F = 2.2$, $p > 0.05$) or of the interaction of SAPS group and task identity ($F = 1.5$, $p > 0.05$) on fair wages.
(TIF)

S3 Fig. Real fair wage values versus simulated fair wage values. For each subject, we simulated data using the model with the highest model responsibility for that subject, and their fit parameter values. Here we have selected 2 random subjects from each NFC tertile (left: low NFC, middle: middle NFC, right: high NFC) and plotted their real and fit fair wage values. The title of each plot is the mean r-squared value after 100 simulations with the subject's best fit model and best fit parameter values. Markers are shaded such that later trials are displayed in darker colors, and the shape of the marker indicates which task was rated (squares are 1-back ratings, circles are 3-detect ratings, and diamonds are 2-back ratings).
(TIF)

S4 Fig. The results of a generate and recover procedure on a model including update, maintenance, and false alarm (FA) costs. A dataset of simulated subjects was produced with random parameter values (constrained by the bounds of those parameters), and then fit with the same procedure as real subject data. Here we show the fits for 50 subjects out of 100, where each subject's fits are plotted in a unique color. The identity line is overlaid on each subplot in black. Comparing the fit parameter values to the real values reveals the high fidelity of the model fitting procedure. Models were fitted using the Computational Behavioral Modeling (cbm) toolbox of Piray et al (2019). All candidate models were visually inspected and verified as recoverable to avoid fitting models with parameter tradeoffs. Only models with parameter recoverability were fit to real subject data.
(TIF)

S1 Text. Covariance between cost components, but not between cost parameters.
(DOCX)

S2 Text. Supplementary model fits and analyses.
(DOCX)

Acknowledgments

We thank our subjects on Amazon Mechanical Turk for their participation in this experiment, as well as our earliest pilot subjects, the members of Arbeitsgruppe Peter Dayan (AGPD) at the Max Planck Institute for Biological Cybernetics. We also thank the members of AGPD for many helpful comments on the initial design of the experiment, and specifically thank Dr. Franziska Broeker for her assistance in establishing the online data collection pipeline, and Drs. Andrew Webb and Aenne Brielmann for thoroughly reviewing the modeling and analysis code for this project. We thank Dr. Martin Breidt, Dr. Holger Dinkel, Mihai Vintiloiu, and the entire IT Core Facility at the Max Planck Campus in Tuebingen for their technological and

research administrative support. We thank Finn van Krieken for his assistance with front-end online experiment development. We also thank Dagmar Maier for administrative support.

Author Contributions

Conceptualization: Sarah L. Master, Peter Dayan.

Formal analysis: Sarah L. Master.

Funding acquisition: Peter Dayan.

Investigation: Sarah L. Master.

Methodology: Sarah L. Master, Peter Dayan.

Project administration: Clayton E. Curtis.

Resources: Clayton E. Curtis, Peter Dayan.

Software: Sarah L. Master.

Supervision: Clayton E. Curtis, Peter Dayan.

Validation: Clayton E. Curtis, Peter Dayan.

Visualization: Sarah L. Master, Clayton E. Curtis, Peter Dayan.

Writing – original draft: Sarah L. Master, Peter Dayan.

Writing – review & editing: Sarah L. Master, Clayton E. Curtis, Peter Dayan.

References

1. Cacioppo JT, Petty RE. The need for cognition. J Pers Soc Psychol. 1982; 42(1):116.
2. Cacioppo JT, Petty RE, Feng Kao C. The efficient assessment of need for cognition. J Pers Assess. 1984; 48(3):306–7. https://doi.org/10.1207/s15327752jpa4803_13 PMID: 16367530
3. Inzlicht M, Shenhav A, Olivola CY. The Effort Paradox: Effort Is Both Costly and Valued. Trends Cogn Sci. 2018 Apr; 22(4):337–49. https://doi.org/10.1016/j.tics.2018.01.007 PMID: 29477776
4. Kool W, McGuire JT, Rosen ZB, Botvinick MM. Decision making and the avoidance of cognitive demand. J Exp Psychol Gen. 2010; 139(4):665. https://doi.org/10.1037/a0020198 PMID: 20853993
5. Kool W, Botvinick M. A labor/leisure tradeoff in cognitive control. Journal of experimental psychology General. 2014; 143(1):131–41. https://doi.org/10.1037/a0031048 PMID: 23230991
6. Sandra DA, Otto AR. Cognitive capacity limitations and Need for Cognition differentially predict reward-induced cognitive effort expenditure. Cognition. 2018 Mar; 172:101–6. https://doi.org/10.1016/j.cognition.2017.12.004 PMID: 29247878
7. Westbrook A, Kester D, Braver TS. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. PLoS One. 2013 Jul 22; 8(7):e68210. https://doi.org/10.1371/journal.pone.0068210 PMID: 23894295
8. Sayalı C, Badre D. Neural systems of cognitive demand avoidance. Neuropsychologia. 2019 Feb 4; 123:41–54. https://doi.org/10.1016/j.neuropsychologia.2018.06.016 PMID: 29944865
9. Sayalı C, Badre D. Neural systems underlying the learning of cognitive effort costs [Internet]. Cold Spring Harbor Laboratory. 2020 [cited 2021 Mar 5]. p. 2020.06.08.139618. Available from: https://www.biorxiv.org/content/10.1101/2020.06.08.139618v1.abstract
10. Agrawal M, Mattar MG, Cohen JD, Daw ND. The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. Psychol Rev. 2022 Apr; 129(3):564–85. https://doi.org/10.1037/rev0000309 PMID: 34383523
11. Constantino SM, Daw ND. Learning the opportunity cost of time in a patch-foraging task. Cogn Affect Behav Neurosci. 2015 Dec; 15(4):837–53. https://doi.org/10.3758/s13415-015-0350-y PMID: 25917000
12. Otto AR, Daw ND. The opportunity cost of time modulates cognitive effort. Neuropsychologia. 2019 Feb 4; 123:92–105. https://doi.org/10.1016/j.neuropsychologia.2018.05.006 PMID: 29750987

13. Callaway F, Jain YR, van Opheusden B, Das P, Iwama G, Gul S, et al. Leveraging artificial intelligence to improve people's planning strategies. Proc Natl Acad Sci U S A. 2022 Mar 22; 119(12): e2117432119. https://doi.org/10.1073/pnas.2117432119 PMID: 35294284

14. Felso V, Lieder F. Measuring individual differences in the depth of planning [Internet]. 2022. Available from: psyarxiv.com/xmf3y

15. Ho MK, Abel D, Correa CG, Littman ML, Cohen JD, Griffiths TL. People construct simplified mental representations to plan. Nature. 2022 Jun; 606(7912):129–36. https://doi.org/10.1038/s41586-022-04743-9 PMID: 35589843

16. Kool W, Gershman SJ, Cushman FA. Planning Complexity Registers as a Cost in Metacontrol. J Cogn Neurosci. 2018 Oct; 30(10):1391–404. https://doi.org/10.1162/jocn_a_01263 PMID: 29668390

17. Sayalı C, Rubin-McGregor J, Badre D. Policy abstraction as a predictor of cognitive effort avoidance. J Exp Psychol Gen. 2023 Dec; 152(12):3440–58. https://doi.org/10.1037/xge0001449 PMID: 37616076

18. Bustamante L, Lieder F, Musslick S, Shenhav A, Cohen J. Learning to Overexert Cognitive Control in a Stroop Task. Cogn Affect Behav Neurosci [Internet]. 2021 Jan 6; Available from: http://dx.doi.org/10.3758/s13415-020-00845-x

19. Cohen JD, Perlstein WM, Braver TS, Nystrom LE, Noll DC, Jonides J, et al. Temporal dynamics of brain activation during a working memory task. Nature. 1997 Apr 10; 386(6625):604–8. https://doi.org/10.1038/386604a0 PMID: 9121583

20. MacLeod CM. The Stroop task: The"gold standard" of attentional measures. J Exp Psychol Gen. 1992; 121(1):12.

21. Becker GM, DeGroot MH, Marschak J. Measuring utility by a single-response sequential method. Behav Sci. 1964 Jul; 9(3):226–32. https://doi.org/10.1002/bs.3830090304 PMID: 5888778

22. De Martino B, Kumaran D, Holt B, Dolan RJ. The neurobiology of reference-dependent value computation. J Neurosci. 2009 Mar 25; 29(12):3833–42. https://doi.org/10.1523/JNEUROSCI.4832-08.2009 PMID: 19321780

23. Boureau YL, Sokol-Hessner P, Daw ND. Deciding How To Decide: Self-Control and Meta-Decision Making. Trends Cogn Sci. 2015 Nov; 19(11):700–10. https://doi.org/10.1016/j.tics.2015.08.013 PMID: 26483151

24. Frömer R, Lin H, Dean Wolf CK, Inzlicht M, Shenhav A. When effort matters: Expectations of reward and efficacy guide cognitive control allocation [Internet]. bioRxiv. 2020 [cited 2022 Aug 11]. p. 2020.05.14.095935. Available from: https://www.biorxiv.org/content/10.1101/2020.05.14.095935

25. Kurzban R, Duckworth A, Kable JW, Myers J. An opportunity cost model of subjective effort and task performance. Behav Brain Sci. 2013 Dec; 36(6):661–79. https://doi.org/10.1017/S0140525X12003196 PMID: 24304775

26. Lieder F, Shenhav A, Musslick S, Griffiths TL. Rational metareasoning and the plasticity of cognitive control. PLoS Comput Biol. 2018 Apr; 14(4):e1006043. https://doi.org/10.1371/journal.pcbi.1006043 PMID: 29694347

27. Shenhav A, Botvinick MM, Cohen JD. The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron. 2013 Jul 24; 79(2):217–40. https://doi.org/10.1016/j.neuron.2013.07.007 PMID: 23889930

28. Dunn TL, Inzlicht M, Risko EF. Anticipating cognitive effort: roles of perceived error-likelihood and time demands. Psychol Res. 2019 Jul; 83(5):1033–56. https://doi.org/10.1007/s00426-017-0943-x PMID: 29134281

29. Shenhav A, Fahey MP, Grahek I. Decomposing the motivation to exert mental effort. Curr Dir Psychol Sci. 2021 Aug 1; 30(4):307–14. https://doi.org/10.1177/09637214211009510 PMID: 34675454

30. Rice KG, Richardson CME, Tueller S. The short form of the revised almost perfect scale. J Pers Assess. 2014; 96(3):368–79. https://doi.org/10.1080/00223891.2013.838172 PMID: 24090303

31. Braver TS, Cohen JD, Nystrom LE, Jonides J, Smith EE, Noll DC. A parametric study of prefrontal cortex involvement in human working memory. Neuroimage. 1997 Jan; 5(1):49–62. https://doi.org/10.1006/nimg.1996.0247 PMID: 9038284

32. Piray P, Dezfouli A, Heskes T, Frank MJ, Daw ND. Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. PLoS Comput Biol. 2019 Jun; 15(6):e1007043. https://doi.org/10.1371/journal.pcbi.1007043 PMID: 31211783

33. Westbrook A, Braver TS. Cognitive effort: A neuroeconomic approach. Cogn Affect Behav Neurosci. 2015 Jun; 15(2):395–415. https://doi.org/10.3758/s13415-015-0334-y PMID: 25673005

34. Milyavskaya M, Inzlicht M, Johnson T, Larson MJ. Reward sensitivity following boredom and cognitive effort: A high-powered neurophysiological investigation. Neuropsychologia. 2019 Feb 4; 123:159–68. https://doi.org/10.1016/j.neuropsychologia.2018.03.033 PMID: 29601888

35. Boag RJ, Stevenson N, van Dooren R, Trutti AC, Sjoerds Z, Forstmann BU. Cognitive Control of Working Memory: A Model-Based Approach. Brain Sci [Internet]. 2021 May 28; 11(6). Available from: https://doi.org/10.3390/brainsci11060721 PMID: 34071635

36. Murray JD, Jaramillo J, Wang XJ. Working Memory and Decision-Making in a Frontoparietal Circuit Model. J Neurosci. 2017 Dec 13; 37(50):12167–86. https://doi.org/10.1523/JNEUROSCI.0343-17.2017 PMID: 29114071

37. Musslick S, Bizyaeva A, Agaron S, Leonard N, Cohen JD. Stability-flexibility dilemma in cognitive control: A dynamical system perspective. Proceedings of the 41st Annual Meeting of the Cognitive Science Society [Internet]. 2019 Jan [cited 2022 Dec 3]; Available from: https://par.nsf.gov/biblio/10125021

38. Baumeister RF, Muraven M, Tice DM. Ego Depletion: A Resource Model of Volition, Self-Regulation, and Controlled Processing. Soc Cogn. 2000 Jun 1; 18(2):130–50.

39. Kurzban R, Duckworth A, Kable JW, Myers J. An opportunity cost model of subjective effort and task performance [Internet]. Vol. 36, Behavioral and Brain Sciences. 2013. p. 661–79. Available from: http://dx.doi.org/10.1017/s0140525x12003196

40. Wiehler A, Branzoli F, Adanyeguh I, Mochel F, Pessiglione M. A neuro-metabolic account of why daylong cognitive work alters the control of economic decisions. Curr Biol. 2022 Aug 22; 32(16):3564–75.e5. https://doi.org/10.1016/j.cub.2022.07.010 PMID: 35961314

41. Blain B, Hollard G, Pessiglione M. Neural mechanisms underlying the impact of daylong cognitive work on economic decisions. Proc Natl Acad Sci U S A. 2016; 113(25):6967–72. https://doi.org/10.1073/pnas.1520527113 PMID: 27274075

42. Mækelæ MJ, Klevjer K, Westbrook A, Eby NS, Eriksen R, Pfuhl G. Is it cognitive effort you measure? Comparing three task paradigms to the Need for Cognition scale. PLoS One. 2023 Aug 17; 18(8):e0290177. https://doi.org/10.1371/journal.pone.0290177 PMID: 37590223

43. Kool W, Botvinick M. A labor/leisure tradeoff in cognitive control. J Exp Psychol Gen. 2014 Feb; 143(1):131–41. https://doi.org/10.1037/a0031048 PMID: 23230991

44. Cools R, D'Esposito M. Inverted-U–Shaped Dopamine Actions on Human Working Memory and Cognitive Control. Biol Psychiatry. 2011 Jun 15; 69(12):e113–25. https://doi.org/10.1016/j.biopsych.2011.03.028 PMID: 21531388

45. Froudist-Walsh S, Bliss DP, Ding X, Rapan L, Niu M, Knoblauch K, et al. A dopamine gradient controls access to distributed working memory in the large-scale monkey cortex. Neuron. 2021 Nov 3; 109(21):3500–20.e13. https://doi.org/10.1016/j.neuron.2021.08.024 PMID: 34536352

46. O'Reilly RC, Frank MJ. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput. 2006 Feb; 18(2):283–328. https://doi.org/10.1162/089976606775093909 PMID: 16378516

47. Lieder F, Griffiths T. Helping people make better decisions using optimal gamification. In: CogSci [Internet]. 2016. Available from: https://re.is.mpg.de/uploads_file/attachment/attachment/610/GamificationCogSciRevised.pdf

48. Lieder F, Krueger PM, Callaway F, Griffiths T. A reward shaping method for promoting metacognitive learning [Internet]. 2017. Available from: psyarxiv.com/qj346

49. Lieder F, Chen OX, Krueger PM, Griffiths TL. Cognitive prostheses for goal achievement. Nat Hum Behav. 2019 Oct; 3(10):1096–106. https://doi.org/10.1038/s41562-019-0672-9 PMID: 31427788

50. Kool W, Botvinick M. Mental labour. Nat Hum Behav. 2018 Dec; 2(12):899–908. https://doi.org/10.1038/s41562-018-0401-9 PMID: 30988433

51. Froböse MI, Westbrook A, Bloemendaal M, Aarts E, Cools R. Catecholaminergic modulation of the cost of cognitive control in healthy older adults. PLoS One. 2020 Feb 21; 15(2):e0229294. https://doi.org/10.1371/journal.pone.0229294 PMID: 32084218

52. Strobel A, Wieder G, Paulus PC, Ott F, Pannasch S, Kiebel SJ, et al. Dispositional cognitive effort investment and behavioral demand avoidance: Are they related? PLoS One. 2020 Oct 14; 15(10):e0239817. https://doi.org/10.1371/journal.pone.0239817 PMID: 33052978

53. Hara K, Adams A, Milland K, Savage S, Callison-Burch C, Bigham JP. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2018. p. 1–14. (CHI '18).

54. de Leeuw JR. jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. Behav Res Methods. 2015 Mar; 47(1):1–12. https://doi.org/10.3758/s13428-014-0458-y PMID: 24683129

55. Rac-Lubashevsky R, Kessler Y. Dissociating working memory updating and automatic updating: The reference-back paradigm. J Exp Psychol Learn Mem Cogn. 2016 Jun; 42(6):951–69. https://doi.org/10.1037/xlm0000219 PMID: 26618910

56. Matlab S. Matlab. The MathWorks, Natick, MA [Internet]. 2012; Available from: https://itb.biologie.hu-berlin.de/~kempter/Teaching/2003_SS/gettingstarted.pdf

57. Berinsky AJ, Margolis MF, Sances MW. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys: Separating the shirkers from the workers? Am J Pol Sci. 2014 Jul; 58(3):739–53.

58. Rac-Lubashevsky R, Kessler Y. Decomposing the n-back task: An individual differences study using the reference-back paradigm. Neuropsychologia. 2016 Sep; 90:190–9. https://doi.org/10.1016/j.neuropsychologia.2016.07.013 PMID: 27425422

59. D'Ardenne K, Eshel N, Luka J, Lenartowicz A, Nystrom LE, Cohen JD. Role of prefrontal cortex and the midbrain dopamine system in working memory updating. Proceedings of the National Academy of Sciences. 2012; 109(49):19900–9. https://doi.org/10.1073/pnas.1116727109 PMID: 23086162

60. Hazy TE, Frank MJ, O'Reilly RC. Banishing the homunculus: making working memory work. Neuroscience. 2006 Apr 28; 139(1):105–18. https://doi.org/10.1016/j.neuroscience.2005.04.067 PMID: 16343792

61. O'Reilly RC. Biologically based computational models of high-level cognition. Science. 2006 Oct 6; 314 (5796):91–4. https://doi.org/10.1126/science.1127242 PMID: 17023651

62. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage. 2009 Jul 15; 46(4):1004–17. https://doi.org/10.1016/j.neuroimage.2009.03.025 PMID: 19306932