

<sup>1</sup> When easy is not preferred: A discounting paradigm to assess load-independent task  
<sup>2</sup> preference

<sup>3</sup> Josephine Zerna<sup>†,1</sup>, Christoph Scheffel<sup>†,1</sup>, Corinna Kührt<sup>1</sup>, & Alexander Strobel<sup>1</sup>

<sup>4</sup> <sup>1</sup> Faculty of Psychology, Dresden University of Technology, 01062 Dresden, Germany

<sup>5</sup> Author Note

<sup>6</sup> The authors made the following contributions. Josephine Zerna: Conceptualization,  
<sup>7</sup> Data curation, Methodology, Funding acquisition, Formal analysis, Investigation, Project  
<sup>8</sup> administration, Software, Visualization, Writing - original draft, Writing - review &  
<sup>9</sup> editing; Christoph Scheffel: Conceptualization, Methodology, Funding acquisition,  
<sup>10</sup> Investigation, Project administration, Software, Writing - review & editing; Corinna Kührt:  
<sup>11</sup> Formal analysis, Writing - review & editing; Alexander Strobel: Conceptualization,  
<sup>12</sup> Resources, Supervision, Funding acquistion, Writing - review & editing. <sup>†</sup> Josephine Zerna  
<sup>13</sup> and Christoph Scheffel contributed equally to this work.

<sup>14</sup> Correspondence concerning this article should be addressed to Josephine Zerna,  
<sup>15</sup> Zellescher Weg 17, 01069 Dresden, Germany. E-mail: josephine.zerna@tu-dresden.de

16

## Abstract

17 When individuals set goals, they consider the subjective value (SV) of the anticipated  
18 reward and the required effort, a trade-off that is of great interest to psychological research.  
19 One approach to quantify the SVs of levels of difficulty of a cognitive task is the Cognitive  
20 Effort Discounting Paradigm by Westbrook and colleagues (2013). However, it fails to  
21 acknowledge the highly individual nature of effort, as it assumes a unidirectional, inverse  
22 relationship between task load and SVs. Therefore, it cannot map differences in effort  
23 perception that arise from traits like Need for Cognition, since individuals who enjoy  
24 effortful cognitive activities likely do not prefer the easiest level. We replicated the analysis  
25 of Westbrook and colleagues with an adapted version, the Cognitive and Affective  
26 Discounting (CAD) Paradigm. It quantifies SVs without assuming that the easiest level is  
27 preferred, thereby enabling the assessment of SVs for tasks without objective order of task  
28 load. Results show that many of the 116 participants preferred a more or the most difficult  
29 level. Variance in SVs was best explained by a declining logistic contrast of the  $n$ -back  
30 levels and by the accuracy of responses, while reaction time as a predictor was highly  
31 volatile depending on the preprocessing pipeline. Participants with higher Need for  
32 Cognition scores perceived higher  $n$ -back levels as less effortful and found them less  
33 aversive. Effects of Need for Cognition on SVs in lower levels did not reach significance, as  
34 group differences only emerged in higher levels. The CAD Paradigm appears to be well  
35 suited for assessing and analysing task preferences independent of the supposed objective  
36 task difficulty.

37       *Keywords:* effort discounting, registered report, specification curve analysis, need for  
38 cognition,  $n$ -back

39       Word count: 7000

40 When easy is not preferred: A discounting paradigm to assess load-independent task  
41 preference

42 **Introduction**

43 In everyday life, effort and reward are closely intertwined<sup>1</sup>. With each decision a  
44 person makes, they have to evaluate whether the effort required to reach a goal is worth  
45 being exerted, given the reward they receive when reaching the goal. A reward is  
46 subjectively more valuable if it is obtained with less effort, so the required effort is used as  
47 a reference point for estimating the reward value<sup>1</sup>. However, the cost of the effort itself is  
48 also subjective, and research has not yet established which function best describes the  
49 relationship between effort and cost<sup>2</sup>. Investigating effort and cost is challenging because  
50 “effort is not a property of the target task alone, but also a function of the individual’s  
51 cognitive capacities, as well as the degree of effort voluntarily mobilized for the task, which  
52 in turn is a function of the individual’s reward sensitivity” (p. 209)<sup>2</sup>.

53 One task that is often used to investigate effort is the *n*-back task, a working memory  
54 task in which a continuous stream of stimuli, e.g. letters, is presented on screen.  
55 Participants indicate via button press whether the current stimulus is the same as *n* stimuli  
56 before, with *n* being the level of difficulty between one and six<sup>3</sup>. The *n*-back task is well  
57 suited to investigate effort because it is an almost continuous manipulation of task load as  
58 has been shown by monotonic increases in error rates, reaction times<sup>4</sup>, and brain activity in  
59 areas associated with working memory<sup>5,6</sup>. However, its reliability measures are mixed, and  
60 associations of *n*-back performance and measures such as executive functioning and fluid  
61 intelligence are often inconsistent<sup>4</sup>.

62 A way to quantify the subjective cost of each *n*-back level has been developed by  
63 Westbrook, Kester, and Braver<sup>7</sup>, called the Cognitive Effort Discounting Paradigm  
64 (COG-ED). First, the participants complete the *n*-back levels to familiarize themselves  
65 with the task. Then, 1-back is compared with each more difficult level by asking the

66 participants to decide between receiving a fixed 2\$ for the more difficult level or the flexible  
67 starting value of 1\$ for 1-back. If they choose the more difficult level, the reward for 1-back  
68 increases by 0.50\$, if they choose 1-back, it decreases by 0.50\$. This is repeated five more  
69 times, with each adjustment of the 1-back reward being half of the previous step, while the  
70 reward for the more difficult level remains fixed at 2\$. The idea is to estimate the point of  
71 subjective equivalence, i.e., the monetary ratio at which both offers are equally preferred<sup>7</sup>.  
72 The subjective value (SV) of each more difficult level is then calculated by dividing the  
73 final reward value of 1-back by the fixed 2\$ reward. Westbrook et al.<sup>7</sup> used these SVs to  
74 investigate inter-individual differences in effort discounting. Younger participants showed  
75 lower effort discounting, i.e., they needed a lower monetary incentive for choosing the more  
76 difficult levels over 1-back.

77 The individual degree of effort discounting in the study by Westbrook et al.<sup>7</sup> was also  
78 associated with the participants' scores in Need for Cognition (NFC), a personality trait  
79 describing an individual's tendency to actively seek out and enjoy effortful cognitive  
80 activities<sup>8</sup>. Westbrook et al.<sup>7</sup> conceptualized NFC as a trait measure of effortful task  
81 engagement, providing a subjective self-report of effort discounting for each participant  
82 which could then be related to the SVs as an objective measure of effort discounting. On  
83 the surface, this association stands to reason, as individuals with higher NFC are more  
84 motivated to mobilize cognitive effort because they perceive it as intrinsically rewarding.  
85 Additionally, it has been shown that individuals avoid cognitive effort only to a certain  
86 degree, possibly to retain a sense of self-control<sup>9</sup>, a trait more prominent in individuals  
87 with high NFC<sup>10–12</sup>. However, the relation of NFC and SVs might be confounded, since  
88 other studies utilizing the COG-ED paradigm found the association of NFC and SVs to  
89 disappear after correcting for performance<sup>13</sup> or found no association of NFC and SVs at  
90 all<sup>14</sup>. On the other hand, task load has been shown to be a better predictor of SVs than  
91 task performance<sup>7,15,16</sup>, so more research is needed to shed light on this issue.

92 With the present study, we alter one fundamental assumption of the original

93 COG-ED paradigm: That the easiest  $n$ -back level has the highest SV. We therefore  
94 adapted the COG-ED paradigm in a way that allows the computation of SVs for different  
95  $n$ -back levels without presuming that all individuals inherently prefer the easiest level.  
96 Since we also aim to establish this paradigm for the assessment of tasks with no objective  
97 task load, e.g., emotion regulation tasks<sup>17</sup>, we call it the Cognitive and Affective  
98 Discounting Paradigm (CAD). In the present study, we validated the CAD paradigm by  
99 conceptually replicating the findings of Westbrook et al.<sup>7</sup>. Additionally, we compared the  
100 effort discounting behavior of participants regarding the  $n$ -back task and an emotion  
101 regulation task. The full results of the latter are published in a second Registered Report<sup>17</sup>.  
102 The COG-ED paradigm has been applied to tasks in different domains before, showing  
103 that SVs across task domains correlate<sup>14</sup>, but these tasks had an objective order of task  
104 load, which is not the case for the choice of emotion regulation strategies or other  
105 paradigms where there is no objective order of task load.

106 Our hypotheses were derived from the results of Westbrook et al.<sup>7</sup>. As a manipulation  
107 check, we hypothesized that with increasing  $n$ -back level the (1a) the signal detection  
108 parameter  $d'$  declines, while (1b) reaction time and (1c) perceived task load increase.  
109 Regarding the associations of task load and effort discounting we hypothesized that (2a)  
110 SVs decline with increasing  $n$ -back level, and (2b) they do so even after controlling for  
111 declining task performance. And finally, we hypothesized that the CAD paradigm can show  
112 inter-individual differences in effort discounting, such that participants with higher NFC  
113 have (3a) lower SVs for 1-back but higher SVs for 2- and 3-back, (3b) lower perceived task  
114 load across all levels, and (3c) higher aversion against 1-back but lower aversion against 2-  
115 and 3-back. Each hypothesis is detailed in the Design Table in the Supplementary Material.

116

## Methods

117 We report how we determined our sample size, all data exclusions (if any), all  
118 manipulations, and all measures in the study<sup>cf. 18</sup>. The paradigm was written and

<sup>119</sup> presented using *Psychopy*<sup>19</sup>. We used *R* (Version 4.2.0)<sup>20</sup> with *R Studio* (Version  
<sup>120</sup> 2022.12.0)<sup>21</sup> with the main packages *papaja* (Version 0.1.1)<sup>22</sup>, *afex* (Version 1.2-1)<sup>23</sup>, and  
<sup>121</sup> *BayesFactor* (Version 0.9.12-4.4)<sup>24</sup> for all our analyses.

<sup>122</sup> **Ethics information**

<sup>123</sup> The study protocol complies with all relevant ethical regulations and was approved  
<sup>124</sup> by the ethics committee of the Technische Universität Dresden (reference number  
<sup>125</sup> SR-EK-50012022). Prior to testing, written informed consent was obtained. Participants  
<sup>126</sup> received 24€ in total or course credit for participation.

<sup>127</sup> **Design**

<sup>128</sup> **CAD Paradigm.** Figure 1 illustrates how different modifications of the COG-ED  
<sup>129</sup> paradigm<sup>7</sup> return SVs that do or do not reflect the true preference of a hypothetical  
<sup>130</sup> participant, who likes 2-back most, 3-back less, and 1-back least (for reasons of clarity  
<sup>131</sup> there are only three levels in the example). The COG-ED paradigm, which compares every  
<sup>132</sup> more difficult level with 1-back sets the SV of 1-back to 1, regardless of the response  
<sup>133</sup> pattern. Adding a comparison of the more difficult levels with each other allows the SVs of  
<sup>134</sup> those two levels to be more differentiated, but leaves the SV of 1-back unchanged. Adding  
<sup>135</sup> those same pairs again, but with the opposite assignment of fixed and flexible level, does  
<sup>136</sup> approach the true preference, but has two disadvantages. First, the SVs are still quite alike  
<sup>137</sup> across levels due to the fact that every more difficult level has only been compared with the  
<sup>138</sup> easiest level, and second, having more task levels than just three would lead to an  
<sup>139</sup> exponential increase in comparisons. Therefore, the solution lies in reducing the number of  
<sup>140</sup> necessary comparisons by presenting only one effort discounting round for each possible  
<sup>141</sup> pair of levels after determining for each pair which level should be fixed and which should  
<sup>142</sup> be flexible. This is determined by presenting each possible pair of levels on screen with the  
<sup>143</sup> question “Would you prefer 1€ for level A or 1€ for level B?”. Participants respond by

clicking the respective on-screen button. Each pair is presented three times, resulting in 18 presented pairs, which are fully randomized in order and in the assignment of which level is on the left or right of the screen. For each pair, the level that was chosen by the participant at least two out of three times will be used as the level with a flexible value, which starts at 1€ and changes in every iteration. The other level in the pair will be set to a fixed value of 2€. Then, the effort discounting sensu Westbrook et al.<sup>7</sup> begins, but with all possible pairs and with the individually determined assignment of fixed and flexible level. The order in which the pairs are presented is fully randomized, and each pair goes through all iteration steps of adding/subtracting 0.50€, 0.25€, 0.13€, 0.06€, 0.03€, 0.02€ to/from the flexible level's reward (each adjustment half of the previous one, rounded to two decimals) before moving on to the next one. This procedure allows to compute SVs based on actual individual preference instead of objective task load. For each pair, the SV of the flexible level is 1, as it was preferred when faced with equal rewards, and the SV of the fixed level is the final reward of the flexible level divided by 2€. Each level's "global" SV is calculated as the mean of this level's SVs from all pairs in which it appeared. If the participant has a clear preference for one level, this level's SV will be 1. If not, then no level's SV will be 1, but each level's SV can still be interpreted as an absolute and relative value, so each participant's effort discounting behaviour can still be quantified. The interpretation of SVs in Westbrook et al.<sup>7</sup> was "The minimum relative reward required for me to choose 1-back over this level". So if the SV of 3-back was 0.6, the participant would need to be rewarded with at least 60 % of what they are being offered for doing 3-back to do 1-back instead, forgoing the higher reward for 3-back. In this study, the SV can be interpreted as "The minimum relative reward required for me to choose any other level over this level". Therefore, an SV of 1 indicates that this level is preferred over all others, while SVs lower than 1 indicate that in at least one pair, a different level was preferred over this one.

[FIGURE 1 HERE]

170       **Study procedure.** Healthy participants aged 18 to 30 years were recruited using

171       the software *ORSEE*<sup>25</sup>. Participants completed the personality questionnaires online and

172       then visited the lab for two sessions one week apart. NFC was assessed using the 16-item

173       short form of the Need for Cognition Scale<sup>26,27</sup>. Responses to each item (e.g., “Thinking is

174       not my idea of fun”, recoded) were recorded on a 7-point Likert scale. The NFC scale

175       shows comparably high internal consistency (Cronbach’s  $\alpha > .80$ )<sup>27,28</sup>. Several other

176       personality questionnaires were used in this study but are the topic of the Registered

177       Report for the second lab session<sup>17</sup>. A full list of measures can be found in our Github

178       repository. In the first session, participants provided informed consent and demographic

179       data before completing the computer-based paradigm. The paradigm started with the

180       n-back levels one to four, presented sequentially with two runs per level, consisting of 64

181       consonants (16 targets, 48 non-targets) per run. The levels were referred to by color

182       (1-back: black, 2-back: red, 3-back: blue, 4-back: green) to avoid anchor effects in the

183       effort discounting procedure. To assess perceived task load, we used the 6-item NASA Task

184       Load Index (NASA-TLX)<sup>29</sup>, where participants evaluate their subjective perception of

185       mental load, physical load, effort, frustration, performance, and time pressure during the

186       task on a 20-point scale. At the end of each level, participants filled out the NASA-TLX on

187       a tablet, plus an item with the same response scale, asking them how aversive they found

188       this n-back level. After the n-back task, participants completed the CAD paradigm on

189       screen and were instructed to do so as realistically as possible, even though the displayed

190       rewards were not paid out on top of their compensation. They were told that one of their

191       choices would be randomly picked for the final run of n-back. However, this data was not

192       analyzed as it only served to incentivise truthful behavior and to stay close to the design of

193       Westbrook et al.<sup>7</sup>. After the CAD paradigm, participants filled out a short questionnaire

194       on the tablet, indicating whether they adhered to the instructions (yes/no) and what the

195       primary motivation for their decisions during the effort discounting procedure was (avoid

196       boredom/relax/avoid effort/seek challenge/other).

197        The second session consisted of an emotion regulation task with negative pictures and  
198    the instruction to suppress facial reactions, detach cognitively from the picture content,  
199    and distract oneself, respectively. The paradigm followed the same structure of task and  
200    effort discounting procedure, but participants could decide which strategy they wanted to  
201    reapply in the last block. Study data was collected and managed using REDCap electronic  
202    data capture tools hosted at Technische Universität Dresden<sup>30,31</sup>.

### 203    Sampling plan

204        Sample size determination was mainly based on the results of the analyses of  
205    Westbrook et al.<sup>7</sup> (see Design Table in the Supplementary Material). The hypothesis that  
206    yielded the largest necessary sample size was a repeated measures ANOVA with  
207    within-between interaction of NFC and *n*-back level influencing SVs. Sample size analysis  
208    with *G\*Power*<sup>32,33</sup> indicated that we should collect data from at least 72 participants,  
209    assuming  $\alpha = .05$  and  $\beta = .95$ . However, the sample size analysis for the hypotheses of the  
210    second lab session revealed a larger necessary sample size of 85 participants to find an  
211    effect of  $d = -0.32$  of emotion regulation on facial muscle activity with  $\alpha = .05$  and  
212     $\beta = .95$ . To account for technical errors, noisy physiological data, or participants who  
213    indicate that they did not follow the instructions, we aimed to collect about 50% more data  
214    sets than necessary,  $N = 120$  in total.

### 215    Analysis plan

216        Data collection and analysis were not performed blind to the conditions of the  
217    experiments. We excluded the data of a participant from all analyses, if the participant  
218    stated that they did not follow the instructions, if the investigator noted that the  
219    participant misunderstood the instructions, or if the participant withdrew their consent.  
220    No data was replaced. The performance measure  $d'$  was computed as the difference of the  
221    *z*-transformed hit rate and the *z*-transformed false alarm rate<sup>34</sup>. Reaction time (RT) data

222 was trimmed by excluding all trials with responses faster than 100 ms, as the relevant  
223 cognitive processes cannot have been completed before<sup>35,36</sup>. Aggregated RT values were  
224 described using the median and the median of absolute deviation (*MAD*) as robust  
225 estimates of center and variability, respectively<sup>37</sup>. Error- and post-error trials were  
226 excluded, because RT in the latter is longer due to more cautious behavior<sup>38,39</sup>. To test our  
227 hypotheses, we performed a series of rmANOVAs and an MLM with orthogonal  
228 sum-to-zero contrasts in order to meaningfully interpret results<sup>40</sup>.

229       *Manipulation check.* Declining performance was investigated by calculating an  
230 rmANOVA with six paired contrasts comparing  $d'$  between two levels of 1- to 4-back at a  
231 time. Another rmANOVA with six paired contrasts was computed to compare the median  
232 RT between two levels of 1- to 4-back at a time. To investigate changes in NASA-TLX  
233 ratings, six rmANOVAs were computed, one for each NASA-TLX subscale, and each with  
234 six paired contrasts comparing the ratings between two levels of 1- to 4-back at a time.

235       *Subjective values.* For each effort discounting round, the SV of the fixed level was  
236 calculated by adding or subtracting the last adjustment of 0.02€ from the last monetary  
237 value of the flexible level, depending on the participant's last choice, and dividing this  
238 value by 2€. This yielded an SV between 0 and 1 for the fixed compared with the flexible  
239 level, while the SV of the flexible level was 1. The closer the SV of the fixed level is to 0,  
240 the stronger the preference for the flexible level. All SVs of each level were averaged to  
241 compute one "global" SV for each level. An rmANOVA with four different contrasts were  
242 computed to investigate the association of SVs and the  $n$ -back levels: Declining linear  
243 (3,1,-1,-3), ascending quadratic (-1,1,1,-1), declining logistic (3,2,-2,-3), and positively  
244 skewed normal (1,2,-1,-2) (Supplementary Figure S1). Depending on whether the linear or  
245 one of the other three contrasts fit the curve best, we applied a linear or nonlinear  
246 multi-level model in the next step, respectively.

247       To determine the influence of task performance on the association of SVs and  $n$ -back

level, we performed MLM. We applied restricted maximum likelihood (REML) to fit the model. As an effect size measure for random effects we first calculated the intraclass correlation (ICC), which displays the proportion of variance that is explained by differences between persons. Second, we estimated a random slopes model of  $n$ -back level (level 1, fixed, and random factor: 0-back, 1-back, 2-back, 3-back) predicting SV nested within subjects. As Mussel et al.<sup>41</sup> could show, participants with high versus low NFC not only have a more shallow decline in performance with higher  $n$ -back levels, but show a demand-specific increase in EEG theta oscillations, which has been associated with mental effort. We controlled for performance, i.e.,  $d'$  (level 1, fixed factor, continuous), median RT (level 1, fixed factor, continuous) in order to eliminate a possible influence of declining performance on SV ratings.

$$SV \sim level + d' + medianRT + (level|subject)$$

Level-1-predictors were centered within cluster as recommended by Enders & Tofghi<sup>42</sup>. By this, the model yields interpretable parameter estimates. If necessary, we adjusted the optimization algorithm to improve model fit. We visually inspected the residuals of the model for evidence to perform model criticism. This was done by excluding all data points with absolute standardized residuals above 3 SD. As effect size measures, we calculated pseudo  $R^2$  for our model and  $f^2$  to estimate the effect of  $n$ -back level according to Lorah<sup>43</sup>.

The association of SVs and NFC was examined with an rmANOVA. We subtracted the SV of 1- from 2-back and 2- from 3-back, yielding two SV difference scores per participant. The sample was divided into participants with low and high NFC using a median split. We then computed an rmANOVA with the within-factor  $n$ -back level and the between-factor NFC group to determine whether there is a main effect of level and/or group, and/or an interaction between level and group on the SV difference scores. Post-hoc tests were computed depending on which effect reached significance at  $p < .01$ . To ensure

the validity of this association, we conducted a specification curve analysis<sup>44</sup>, which included 63 possible preprocessing pipelines of the RT data. These pipelines specify which transformation was applied (none, log, inverse, or square-root), which outliers were excluded (none, 2, 2.5, or 3 *MAD* from the median, RTs below 100 or 200 ms), and across which dimensions the transformations and exclusions were applied (across/within subjects and across/within *n*-back levels). The rmANOVA was run with each of the 63 pipelines, which also included our main pipeline (untransformed data, exclusion of RTs below 100 ms). The ratio of pipelines that lead to significant versus non-significant effects provides an indication of how robust the effect actually is.

The association of subjective task load with NFC was examined similarly. We calculated NASA-TLX sum scores per participant per level, computed an rmANOVA with the within-factor *n*-back level and the between-factor NFC group, and applied post-hoc tests based on which effect reached significance at  $p < .01$ . And the association of subjective aversiveness of the task with NFC was examined with difference scores as well, since we expected this curve to mirror the SV curve, i.e. as the SV rises, the aversiveness declines, and vice versa. We subtracted the aversiveness ratings of 1- from 2-back and 2- from 3-back, yielding two aversiveness difference scores per participant. Then, we computed an rmANOVA with the within-factor *n*-back level and the between-factor NFC group, and applied post-hoc tests based on which effect reached significance at  $p < .01$ .

The results of each analysis was assessed on the basis of both *p*-value and the Bayes factor  $BF_{10}$ , calculated with the *BayesFactor* package<sup>24</sup> using the default prior widths of the functions *anovaBF*, *lmBF* and *ttestBF*. We considered a  $BF_{10}$  close to or above 3/10 as moderate/strong evidence for the alternative hypothesis, and a  $BF_{10}$  close to or below .33/.10 as moderate/strong evidence for the null hypothesis<sup>45</sup>.

**296 Pilot data**

297 The sample of the pilot study consisted of  $N = 15$  participants (53.3% female,  
298  $M = 24.43$  ( $SD = 3.59$ ) years old). One participant's data was removed because they  
299 misunderstood the instruction. Due to a technical error the subjective task load data of  
300 one participant was incomplete, so the hypotheses involving the NASA-TLX were analyzed  
301 with  $n = 14$  data sets. The results showed increases in subjective and objective task load  
302 measures with higher  $n$ -back level. Importantly, SVs were lower for higher  $n$ -back levels,  
303 but not different between 1- and 2-back, which shows that the easiest level is not  
304 universally preferred. The MLM revealed  $n$ -back level as a reliable predictor of SV, even  
305 after controlling for declining task performance ( $d'$  and median RT). NASA-TLX scores  
306 were higher with higher  $n$ , and lower for the group with lower NFC scores, but NFC and  
307  $n$ -back level did not interact. All results are detailed in the Supplementary Material.

**308 Data availability**

309 The data of this study can be downloaded from [osf.io/vnj8x/](https://osf.io/vnj8x/).

**310 Code availability**

311 The paradigm code, the R script for analysis, and the R Markdown file used to  
312 compile this document are available at [osf.io/vnj8x/](https://osf.io/vnj8x/).

**313 Protocol registration**

314 The Stage 1 Registered Report protocol has been approved and is available at  
315 [osf.io/cpxth/](https://osf.io/cpxth/).

## Results

### 317 **Adjustments for Stage 2**

318 There were three necessary adjustments of the methods. First, we failed to update  
319 the necessary sample size after the analyses changed with the first review round. Instead of  
320 the 72 subjects stated above, the largest minimum sample size was actually 53 subjects  
321 (see hypothesis 1b in the Design Table in the Supplementary Material). Secondly, we  
322 changed to which hypothesis we applied the specification curve analysis (SCA). In the  
323 initial Stage 1 submission, we had applied it to the MLM of hypothesis 2b, which at this  
324 point included NFC as a predictor. Following the advice of the reviewers, we removed NFC  
325 from the MLM, and analyzed NFC in an rmANOVA (hypothesis 3a) instead. Since NFC  
326 was of great interest to us, we decided to apply the SCA to hypothesis 3a rather than 2b to  
327 provide a measure of robustness. However, hypothesis 3a does not contain any RT data, so  
328 the SCA is only useful for the MLM in hypothesis 2b. Therefore, we applied it to the MLM.  
329 The final adjustment was made during the Stage 2 revision. A fellow researcher made us  
330 aware that by using the z-transformed hit and false alarm rates for the computation of  $d'$ ,  
331 the mean of  $d'$  would be approximately 0 for each  $n$ -back level by design. Consequently,  $d'$   
332 could not show changes across  $n$ -back levels in the manipulation check and would likely  
333 yield different results in the MLM. Therefore, we computed  $d'$  with unstandardized hit and  
334 false alarm rates. We would like to thank Georgia Clay for pointing us to this fallacy.

### 335 **Sample**

336 Data was collected between the 16th of August 2022 and the 3rd of February 2023.  
337 Of the  $N = 176$  participants who filled out the NFC questionnaire,  $n = 124$  completed the  
338 first lab session. Based on the experimenters' notes, we excluded the data of seven  
339 participants from analysis for misunderstanding the instruction of the  $n$ -back task, and the  
340 data of one participant who reported that they confused the colours of the levels during

<sup>341</sup> effort discounting. Our final data set therefore included  $N = 116$  participants (83.60%  
<sup>342</sup> female,  $M \pm SD = 22.4 \pm 3$ ) years old), which is 2.2 times more than what the highest  
<sup>343</sup> sample size calculation required.

<sup>344</sup> **Manipulation checks**

<sup>345</sup> We used rmANOVAs to investigate whether objective performance measures and  
<sup>346</sup> subjective task load measures changed across  $n$ -back levels. For each rmANOVA we report  
<sup>347</sup> the generalized eta squared  $\hat{\eta}_G^2$ , which estimates the effect size in analyses that contain  
<sup>348</sup> both manipulated and non-manipulated terms. In line with hypothesis H1a, the  
<sup>349</sup> performance measure  $d'$  decreased across  $n$ -back levels ( $F(2.74, 315.51) = 197.18, p < .001$ ,  
<sup>350</sup>  $\hat{\eta}_G^2 = .464$ , 90% CI [.403, .516],  $BF_{10} = 1.56 \times 10^{101}$ ). It decreased more strongly from 2- to  
<sup>351</sup> 3-back ( $t(345) = 10.41, p_{Tukey(4)} < .001, BF_{10} = 3.71 \times 10^{23}$ ) than from 1- to 2-back  
<sup>352</sup> ( $t(345) = 4.31, p_{Tukey(4)} < .001, BF_{10} = 93, 954.04$ ) or 3- to 4-back ( $t(345) = 7.18,$   
<sup>353</sup>  $p_{Tukey(4)} < .001, BF_{10} = 2.45 \times 10^{14}$ ). Similarly, the median RT increased across  $n$ -back  
<sup>354</sup> levels ( $F(2.46, 283.05) = 98.67, p < .001, \hat{\eta}_G^2 = .192$ , 90% CI [.130, .248],  
<sup>355</sup>  $BF_{10} = 2.28 \times 10^{34}$ ), supporting hypothesis H1b. Specifically, the median RT was higher  
<sup>356</sup> for the more difficult level in every contrast, with two exceptions: It did not differ between  
<sup>357</sup> 2- and 4-back, and it was higher for 3- than for 4-back (Table 1).

Table 1

*Paired contrasts for the rmANOVA comparing the median reaction time between  $n$ -back levels*

| Contrast | Estimate | SE   | df     | t      | p     | $BF_{10}$             | $\eta_p^2$ | 95%CI        |
|----------|----------|------|--------|--------|-------|-----------------------|------------|--------------|
| 1 - 2    | -0.11    | 0.01 | 345.00 | -11.76 | <.001 | $1.75 \times 10^{30}$ | 0.29       | [0.22, 1.00] |
| 1 - 3    | -0.16    | 0.01 | 345.00 | -16.23 | <.001 | $8.80 \times 10^{45}$ | 0.43       | [0.37, 1.00] |
| 1 - 4    | -0.12    | 0.01 | 345.00 | -12.47 | <.001 | $4.79 \times 10^{34}$ | 0.31       | [0.25, 1.00] |
| 2 - 3    | -0.04    | 0.01 | 345.00 | -4.47  | <.001 | 5,538.45              | 0.05       | [0.02, 1.00] |
| 2 - 4    | -0.01    | 0.01 | 345.00 | -0.71  | 0.894 | 0.10                  | 1.45e-03   | [0.00, 1.00] |
| 3 - 4    | 0.04     | 0.01 | 345.00 | 3.76   | 0.001 | $6.35 \times 10^6$    | 0.04       | [0.01, 1.00] |

*Note.* The column Contrast contains the  $n$  of the  $n$ -back levels. SE = standard error, df = degrees of freedom, t = t-statistic, p = p-value, CI = confidence interval.

All NASA-TLX subscale scores increased across  $n$ -back levels, so evidence was in favour of H1c. Ratings on the effort subscale ( $F(2.20, 253.06) = 203.82, p < .001$ ,  $\hat{\eta}_G^2 = .316, 90\% \text{ CI } [.250, .375]$ ,  $\text{BF}_{10} = 2.47 \times 10^{34}$ ) increased across all levels, but the magnitude of change decreased from 1- to 2-back ( $t(345) = -12.35, p_{\text{Tukey}(4)} < .001$ ,  $\text{BF}_{10} = 4.24 \times 10^{19}$ ) to 3- to 4-back ( $t(345) = -2.72, p_{\text{Tukey}(4)} = .035, \text{BF}_{10} = 174.38$ ). Three subscales had significant differences between all contrasts except for 3- versus 4-back: While ratings on the frustration and time subscales were higher for more difficult levels ( $F(2.50, 287.66) = 68.06, p < .001, \hat{\eta}_G^2 = .172, 90\% \text{ CI } [.112, .227]$ ,  $\text{BF}_{10} = 5.26 \times 10^{15}$ , and  $F(2.21, 254.65) = 51.08, p < .001, \hat{\eta}_G^2 = .117, 90\% \text{ CI } [.065, .168]$ ,  $\text{BF}_{10} = 3.94 \times 10^9$ , respectively), ratings on the performance subscale decreased with higher  $n$  ( $F(2.49, 285.97) = 95.33, p < .001, \hat{\eta}_G^2 = .241, 90\% \text{ CI } [.176, .299]$ ,  $\text{BF}_{10} = 1.55 \times 10^{24}$ ). Ratings on the mental subscale consistently increased across all levels ( $F(1.99, 228.35) = 274.47, p < .001, \hat{\eta}_G^2 = .375, 90\% \text{ CI } [.309, .432]$ ,  $\text{BF}_{10} = 1.64 \times 10^{43}$ ). Ratings on the physical subscale were higher for more difficult levels ( $F(1.68, 192.93) = 15.91, p < .001, \hat{\eta}_G^2 = .041, 90\% \text{ CI } [.009, .075]$ ,  $\text{BF}_{10} = 60.54$ ), apart from the contrasts 2- versus 3-back ( $\text{BF}_{10} = 10.45$ ) and 3- versus 4-back ( $\text{BF}_{10} = 0.47$ ). The full results of these manipulation checks are listed in Table S.1 to S.8 in the Supplementary Material.

### **376 Decline of subjective values**

The different curves of SVs across  $n$ -back levels can be seen in Figure 2, grouped into those participants who had an SV of 1.0 for 1-back ( $n = 71$ ), for 2-back ( $n = 18$ ), for 3-back ( $n = 9$ ), for 4-back ( $n = 13$ ), or all SVs below 1.0, i.e. no absolute preference for any level ( $n = 5$ ). While the majority of participants preferred the easiest level and showed an approximately linear decline of SVs with increasing task-load, a substantial part of the sample had higher SVs for one of the more difficult  $n$ -back levels. However, each panel in Figure 2 contains curves of participants who had large differences between their four SVs

<sup>384</sup> and curves of participants who had a difference of less than 0.2 between their highest and  
<sup>385</sup> their lowest SV, so preferring one level does not necessarily mean having a strong aversion  
<sup>386</sup> against the others, regardless of difficulty level.

<sup>387</sup> [FIGURE 2 HERE]

<sup>388</sup> When asking participants what motivated their decisions in the cognitive effort  
<sup>389</sup> discounting paradigm, 11.2% stated that they wanted to avoid boredom, 22.4% stated that  
<sup>390</sup> they wanted a challenge, 34.5% stated that they wanted to avoid effort, and 4.3% stated  
<sup>391</sup> that they wanted to relax. The remaining 27.6% of participants used the free text field and  
<sup>392</sup> provided reasons such as “I wanted a fair relation of effort and reward.”, “I wanted the fun  
<sup>393</sup> that I had in the more challenging levels.”, “I wanted to maximize reward first and  
<sup>394</sup> minimize effort second.”, or “I did not want to perform poorly when I was being paid for  
<sup>395</sup> it.”. Figure 3 shows the different motivations in the context of the SVs per *n*-back level.

<sup>396</sup> [FIGURE 3 HERE]

<sup>397</sup> The rmANOVA showed a significant difference between the SVs across *n*-back levels  
<sup>398</sup> ( $F(1.98, 227.98) = 65.65, p < .001, \eta^2_G = .288, 90\% \text{ CI } [.222, .347], \text{BF}_{10} = 1.58 \times 10^{64}$ ), so  
<sup>399</sup> evidence was in favour of H2a. All four pre-defined contrasts reached significance (Table 2),  
<sup>400</sup> so a purely linear contrast can be rejected.

Table 2  
*Contrasts for the rmANOVA comparing the subjective values between n-back levels*

| Contrast                 | Estimate | SE   | df     | t     | p     | $\eta_p^2$ | 95%CI        |
|--------------------------|----------|------|--------|-------|-------|------------|--------------|
| Declining Linear         | 1.11     | 0.08 | 345.00 | 13.41 | <.001 | 0.34       | [0.28, 1.00] |
| Ascending Quadratic      | 0.15     | 0.04 | 345.00 | 4.14  | <.001 | 0.05       | [0.02, 1.00] |
| Declining Logistic       | 1.22     | 0.09 | 345.00 | 12.97 | <.001 | 0.33       | [0.26, 1.00] |
| Positively Skewed Normal | 0.75     | 0.06 | 345.00 | 12.74 | <.001 | 0.32       | [0.26, 1.00] |

*Note.* SE = standard error, df = degrees of freedom, t = t-statistic, p = p-value, CI = confidence interval.

<sup>401</sup> The declining logistic contrast had the highest effect estimate ( $t(345) = 12.97$ ,

<sup>402</sup>  $p < .001$ ), suggesting a shallow decline of SVs between 1- and 2-back, and 3- and 4-back,  
<sup>403</sup> respectively, and a steeper decline of SVs between 2- and 3-back. Based on the effect  
<sup>404</sup> estimate, the ascending quadratic and the skewed normal contrasts were rejected in favour  
<sup>405</sup> of the declining logistic contrast.

<sup>406</sup> Consequently, we had to adapt the MLM to incorporate this non-linear trend. To  
<sup>407</sup> apply the contrast to the  $n$ -back levels, we had to turn the variables into a factor, with two  
<sup>408</sup> consequences: Centered variables cannot be turned into factors, so we entered the variable  
<sup>409</sup> level in its raw form, and factors cannot be used as random slopes, so the model is now  
<sup>410</sup> defined as:

$$SV \sim level + d' + medianRT + (1|subject)$$

<sup>411</sup> This means that the intercept still varied between subjects, but there were no random  
<sup>412</sup> slopes anymore. To provide more than one observation per factor level, we used the two  
<sup>413</sup> rounds per  $n$ -back level per subject, rather than  $n$ -back levels per subject. The ICC of the  
<sup>414</sup> null model indicated that there was a correlation of  $r = .096$  between the SVs of a subject,  
<sup>415</sup> i.e. that 9.59% of variance in SVs could be explained by differences between participants.  
<sup>416</sup> We did not use an optimization algorithm to improve the fit of the random intercept  
<sup>417</sup> model. A total of 10 data points from 6 participants were excluded, because the residuals  
<sup>418</sup> exceeded 3 SD above the mean. The results of the final model are displayed in Table 3.

Table 3

*Results of the multi level model on the influence of n-back level (as a declining logistic contrast) and task performance on subjective values.*

| Parameter    | Beta | SE   | df     | t-value | p-value | $f^2$ | Random Effects (SD) |
|--------------|------|------|--------|---------|---------|-------|---------------------|
| Intercept    | 0.81 | 0.01 | 115.00 | 78.68   | <.001   |       | 0.09                |
| n-back level | 0.03 | 0.00 | 797.54 | 9.99    | <.001   | 0.21  |                     |
| d'           | 0.21 | 0.03 | 797.63 | 6.28    | <.001   | 0.05  |                     |
| median RT    | 0.03 | 0.07 | 797.80 | 0.42    | 0.674   | 0.00  |                     |

*Note.* SE = standard error, df = degrees of freedom, SD = standard deviation.

An exploratory ANOVA was used to compare the fit of the final model with a linear random intercept model, confirming that the two models were different from each other ( $\chi^2(2) = 28.35, p < .001$ ), and with an Akaike Information Criterion of  $AIC = -510.93$  and a Bayesian Information Criterion of  $BIC = -472.35$  the declining logistic model was superior to the linear model ( $AIC = -486.58, BIC = -457.65$ ). Both AIC and BIC subtract the likelihood of the model from the number of parameters and/or data points, so lower values indicate better model fit. The final model had an effect size of  $f^2 = 0.21$  for the  $n$ -back levels and  $f^2 = 0.05$  for  $d'$ , which are considered medium and small, respectively<sup>46</sup>. This means that the  $n$ -back level explained 20.67% and  $d'$  explained 4.90% of variance in SVs relative to the unexplained variance, respectively. The beta coefficient indicated that with every 1-unit increase in  $d'$ , the SV increased by 0.21. Due to the coding scheme of the logistic contrast, the beta coefficient of the  $n$ -back level has to be interpreted inversely, so SVs decline with increasing  $n$ -back level. The effect size of the median RT was  $f^2 = 0.00$ . Since SVs decline with increasing level, beyond the variance explained by  $d'$ , evidence was in favour of H2b.

To investigate the dependency of the model results on the RT preprocessing, we conducted a specification curve analysis (Figure 4).

[FIGURE 4 HERE]

Regardless of the preprocessing pipeline,  $n$ -back level and  $d'$  were significant predictors of SVs, and had stable effect estimates across all pipelines. There was no pipeline in which the median RT was a significant predictor of SVs, but it showed volatile effect estimates across pipelines. The pipelines that yielded the highest estimates for  $d'$  and the median RT used log-transformed data, irrespective of the dimension and exclusion criteria.

<sup>442</sup> **Differences between NFC groups**

<sup>443</sup> The median NFC was 16, with  $n = 57$  subjects below and  $n = 59$  above the median.  
<sup>444</sup> We used an rmANOVA to investigate whether the difference between the SVs of 1- and  
<sup>445</sup> 2-back, and 2- and 3-back, respectively, depended on whether a participant's NFC score  
<sup>446</sup> was above or below the median. There was a main effect of the  $n$ -back level  
<sup>447</sup> ( $F(1, 114) = 9.13, p = .003, \hat{\eta}_G^2 = .040, 90\% \text{ CI } [.002, .115], \text{BF}_{10} = 12.68 \pm 0.00\%$ ), but  
<sup>448</sup> neither a main effect of the NFC group ( $F(1, 114) = 3.18, p = .077, \hat{\eta}_G^2 = .013, 90\% \text{ CI }$   
<sup>449</sup> [.000, .068],  $\text{BF}_{10} = 0.56 \pm 0.03\%$ ) nor an interaction of NFC group and  $n$ -back level  
<sup>450</sup> ( $F(1, 114) = 0.46, p = .499, \hat{\eta}_G^2 = .002, 90\% \text{ CI } [.000, .037]$ ), so evidence was not in favour  
<sup>451</sup> of H3a. Post-hoc tests showed that the difference between the SVs of 2- and 3-back is  
<sup>452</sup> slightly more negative than the difference between 1- and 2-back ( $t(114) = -3.02,$   
<sup>453</sup>  $p = .003$ ), but there were large inter-individual differences (Supplementary Figure S2a).  
<sup>454</sup> This means that across the whole sample, there was a steeper decline in SVs from 2- to  
<sup>455</sup> 3-back than from 1- to 2-back, again resembling the declining logistic function.

<sup>456</sup> The rmANOVA on the association between NFC scores and NASA-TLX scores  
<sup>457</sup> revealed a main effect of  $n$ -back level ( $F(2.10, 239.56) = 154.50, p < .001, \hat{\eta}_G^2 = .223, 90\%$   
<sup>458</sup>  $\text{CI } [.159, .282], \text{BF}_{10} = 2.22 \times 10^{45}$ ) and an interaction between  $n$ -back level and NFC scores  
<sup>459</sup> ( $F(2.10, 239.56) = 4.93, p = .007, \hat{\eta}_G^2 = .009, 90\% \text{ CI } [.000, .025]$ ), but no main effect of  
<sup>460</sup> NFC scores ( $F(1, 114) = 3.22, p = .075, \hat{\eta}_G^2 = .022, 90\% \text{ CI } [.000, .084], \text{BF}_{10} = 1.75 \times 10^2$ ).  
<sup>461</sup> Post-hoc tests showed that the participants with NFC scores below the median had higher  
<sup>462</sup> NASA-TLX scores for 3-back ( $t(114) = -2.15, p = .033, \text{BF}_{10} = 11.15$ ) and for 4-back  
<sup>463</sup> ( $t(114) = -2.89, p = .005, \text{BF}_{10} = 336.88$ ) than those with NFC scores above the median,  
<sup>464</sup> so evidence was in favour of H3b. Regardless of NFC scores, NASA-TLX scores were  
<sup>465</sup> higher for the more difficult level in each pair of  $n$ -back levels (Supplementary Figure S3).

<sup>466</sup> With another rmANOVA we investigated whether the difference between the  
<sup>467</sup> aversiveness scores of 1- and 2-back, and 2- and 3-back, respectively, depended on whether

468 a participant's NFC score was above or below the median. There was a main effect of NFC  
 469 group ( $F(1, 114) = 8.43, p = .004, \hat{\eta}_G^2 = .043, 90\% \text{ CI } [.003, .119], \text{BF}_{10} = 14.26$ ) and a  
 470 main effect of the  $n$ -back level ( $F(1, 114) = 10.21, p = .002, \hat{\eta}_G^2 = .034, 90\% \text{ CI } [.000, .105]$ ,  
 471 ), but no interaction ( $F(1, 114) = 2.59, p = .110, \hat{\eta}_G^2 = .009, 90\% \text{ CI } [.000, .058]$ ). In favour  
 472 of H3c, post-hoc tests revealed that participants with NFC scores below the median  
 473 reported higher aversiveness than participants with NFC scores above the median  
 474 ( $t(114) = 2.90, p = .004$ ) (Supplementary Figure S2b). Regardless of NFC, the difference of  
 475 the aversiveness scores of 2- and 3-back was more negative than that of 1- and 2-back  
 476 ( $t(114) = 3.20, p = .002$ ), indicating that in the same way in which the SVs decreased more  
 477 strongly from 2- to 3-back than from 1- to 2-back, the aversion increased more strongly.  
 478 The full results of these analyses of NFC group differences can be found in Table S.11 to  
 479 S.15 in the Supplementary Material.

#### 480 Exploratory analyses

481 To investigate the apparent group difference between the SVs of participants with  
 482 NFC scores below and above the median in higher  $n$ -back levels, we computed an  
 483 rmANOVA with the within-factor level (1 to 4) and the between-factor NFC group  
 484 (below/above median). There was no main effect of NFC group ( $F(1, 114) = 2.63,$   
 485  $p = .108, \hat{\eta}_G^2 = .007, 90\% \text{ CI } [.000, .053], 2.95 \times 10^{-1}$ ), but a main effect of the  $n$ -back level  
 486 ( $F(2.01, 229.39) = 67.39, p < .001, \hat{\eta}_G^2 = .295, 90\% \text{ CI } [.228, .354], 2.70 \times 10^{30}$ ) and an  
 487 interaction ( $F(2.01, 229.39) = 3.24, p = .041, \hat{\eta}_G^2 = .020, 90\% \text{ CI } [.000, .044]$ ). Post-hoc  
 488 tests for the main effect of level showed that SVs were lower for the more difficult  $n$ -back  
 489 level in each paired contrast except for 1- versus 2-back. Post-hoc tests for the interaction  
 490 effect showed that the NFC groups only had a significant difference in SVs for 4-back,  
 491 where participants below the NFC median had lower scores ( $\Delta M = 0.11, 95\% \text{ CI }$   
 492  $[0.01, 0.22], t(114) = 2.13, p = .036$ ). Despite not reaching significance, 1-back was the only  
 493 level in which participants with NFC scores above the median seemed to have lower SVs

494 than those with scores below the median ( $\Delta M = -0.05$ , 95% CI  $[-0.11, 0.01]$ ,  
495  $t(114) = -1.50$ ,  $p = .136$ ). The full results of this exploratory analysis of NFC group  
496 differences can be found in Table S.16 and S.17 in the Supplementary Material.  
497 Supplementary Figure S4 shows the SVs per  $n$ -back level for participants with NFC scores  
498 above and below the median.

499 Following a reviewer's recommendation, we also analyzed the association of SVs with  
500 NFC as a continuous variable. We computed an rmANOVA with the  $n$ -back level as a  
501 within variable and the standardized NFC score as a covariate to predict SVs. Both the  
502 NFC score ( $F(1, 114) = 4.34$ ,  $p = .039$ ,  $\hat{\eta}_G^2 = .011$ , 90% CI [.000, .063],  $BF_{10} = 0.57$ ) and  
503 the  $n$ -back level ( $F(2.02, 229.75) = 67.24$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .295$ , 90% CI [.228, .354],  
504  $BF_{10} = 2.70 \times 10^{30}$ ) showed significant main effects, as well as a significant interaction  
505 ( $F(2.02, 229.75) = 3.78$ ,  $p = .024$ ,  $\hat{\eta}_G^2 = .023$ , 90% CI [.000, .049],  $BF_{10} = 0.12$ ). Analyzing  
506 the estimated marginal means of the linear trends for each  $n$ -back level indicated a  
507 significant difference between the slopes of 1-back and 4-back ( $\Delta M = -0.09$ , 95% CI<sub>Tukey(4)</sub>  
508  $[-0.15, -0.02]$ ,  $t(456) = -3.22$ ,  $p_{Tukey(4)} = .008$ ), but not between any other two levels.  
509 Plotting the predicted slopes shows that there is a negative association between the  
510 predicted SVs and the NFC scores for 1-back, but a positive association between the  
511 predicted SVs and the NFC scores for 4-back (Figure 5). The full results of this  
512 exploratory analysis of NFC as a continuous covariate can be found in Table S.18 and S.19  
513 in the Supplementary Material.

514 [FIGURE 5 HERE]

515

## Discussion

516 This Registered Report aimed to adapt the Cognitive Effort Discounting (COG-ED)  
517 paradigm by Westbrook et al.<sup>7</sup>, which estimates subjective values of different  $n$ -back levels,  
518 into the Cognitive and Affective Discounting (CAD) paradigm to estimate SVs of tasks

519 without defaulting to the assumed objective task load as a benchmark. For this purpose,  
520 we adapted the way in which the discounting options are presented to the participants,  
521 based the anchor on their own choices, and computed SVs across multiple combinations of  
522 task levels. The analyses were closely aligned with those in Westbrook et al.<sup>7</sup> to  
523 demonstrate the changes in SVs brought about by the new paradigm. This study also  
524 applied the CAD paradigm to an emotion regulation task, the results of which are detailed  
525 in a second Registered Report<sup>17</sup>.

## 526 Manipulation checks

527 Both  $d'$  and the median RT changed across  $n$ -back levels, indicating that there was  
528 an increase in objective task load. The steepest decrease in  $d'$  between levels was from 2- to  
529 3-back, resembling the declining logistic curve of the SVs. Interestingly, the median RT  
530 increased from 2- to 3-back, but did not differ between 2- and 4-back. Since feedback was  
531 given after each round, this pattern could be interpreted in such a way that participants  
532 tried to compensate their lower accuracy in 3-back by relying more on luck than on  
533 memory in 4-back and prioritized speed over accuracy. Furthermore, several participants  
534 said that they perceived 3-back as more difficult than 4-back because they found it is easier  
535 to remember chunks of stimuli when  $n$  was an even number than when  $n$  was an odd  
536 number. This would support the notion that the manipulation of task load in an  $n$ -back  
537 task is not strictly continuous. And lastly, the fact that neither accuracy nor speed is an  
538 informative performance measure by itself has been observed before<sup>47</sup> and both show  
539 different associations with various measures of intelligence<sup>4</sup>, suggesting that they should  
540 always be reported as separate indices.

541 All NASA-TLX subscales differed across  $n$ -back levels, but the effort and mental load  
542 subscales were the only ones to consistently increase across all levels. This would support  
543 the notion of the  $n$ -back task offering a continuous manipulation of task load, at least  
544 subjectively. Ratings on the frustration and time subscales increased and ratings on the

545 performance subscale decreased until 3-back and then remained stable. This pattern is  
546 akin to the RT, which also increased and then remained stable. Ratings on the physical  
547 load subscale increased with  $n$ -back levels, but not between 2- and 3-back and 3- and  
548 4-back, respectively.

549 **Decline of subjective values**

550 The rmANOVA with different pre-defined contrasts showed that all fit the SVs to a  
551 different degree, and that the SVs do not simply decline linearly across  $n$ -back levels. The  
552 best fit was a declining logistic curve, reflecting that 1) the majority of participants  
553 preferred the easiest level, 2) 2-back was generally closer in preference to 1-back than  
554 3-back was to 2-back, and 3) objective task load and subjective preference do not stand in  
555 a linear relationship. Since the majority of participants preferred the easiest level, we  
556 rejected the ascending quadratic and skewed normal contrasts, which implied lower SVs for  
557 1- than for 2-back. The fact that the majority of participants preferred lower over higher  
558 effort, but a minority showed the opposite pattern, is in line with previous research on  
559 cognitive effort by Kool et al. (2010)<sup>48</sup>. Importantly, having a paradigm that can  
560 accurately assess the preferences of the minority is necessary but not sufficient, because the  
561 interindividual variability is so high that it blurs effects on the group level. Figure 2  
562 suggests that those who prefer either 1- or 2-back have a slightly steeper discounting curve  
563 than those who prefer 3- or 4-back, meaning they have lower SVs for higher levels than  
564 those who prefer a higher level have for the easier levels. But as the figure also shows, there  
565 is great interindividual variability in the discounting patterns, regardless of which level has  
566 the highest SV. Thomson and Oppenheimer<sup>49</sup> argue that the different effort curves that  
567 have been observed for different tasks are likely due to the fact that we still understand  
568 quite little about how and why different manipulations of effort work. For example, the  
569  $n$ -back task is likely not a continuous manipulation of task load, as discussed above.  
570 However, the declining logistic curve is similar to the sigmoidal curve that has been found

571 for a physical<sup>50</sup> and a cognitive effort paradigm<sup>51</sup>, suggesting there are common features of  
572 effort across different tasks and domains. The MLM with the logistic contrast showed that  
573 the  $n$ -back level explained the majority of variance in SVs, while the performance measure  
574  $d'$  also explained some variance in SVs, albeit less. With increasing  $n$ -back level and  
575 decreasing  $d'$ , the SV decreased. The median RT was not a significant predictor in this  
576 model. Participants might have deliberately or subconsciously used the feedback they  
577 received at the end of each round, i.e. twice per  $n$ -back level, as an anchor during the effort  
578 discounting. This feedback was based on correct responses and not on RT, so if  
579 participants based their effort discounting choices at least partly on this feedback, they  
580 were either motivated to repeat a task in which they performed well and/or they were  
581 reluctant to accept a larger reward for a task in which they did not perform well. Since  
582 more participants reported effort avoidance as their motivation in the effort discounting  
583 than those who reported seeking a challenge, we can assume that they were more  
584 motivated to repeat a task in which they performed well because their good performance  
585 coincided with low effort.

586 The declining logistic  $n$ -back levels and  $d'$  remained significant predictors of SVs  
587 throughout all 63 preprocessing pipelines in the specification curve analysis. The effect  
588 estimates of the  $n$ -back level varied by about 0.00, those of  $d'$  by about 0.07. In contrast to  
589 this stood the variability of the median RT effect estimates (around 0.16), which did not  
590 reach significance in any pipeline. Interestingly, the curve of median RT betas in Figure 4a  
591 roughly mirrored the rectangular pipeline indicators in the transformation rows of  
592 Figure 4b, so the transformation choice influenced the median RT much more than the  
593 dimension or the exclusion choice did. As Fernandez et al.<sup>52</sup> found, applying more than one  
594 preprocessing step to the reaction time data of a Stroop task increased the risk of false  
595 positives beyond  $\alpha = .05$ , and transformation choices inflated this risk more than outlier  
596 exclusion or aggregation choices did. Our data seems to corroborate this finding for  $n$ -back  
597 tasks as well. Surprisingly, the  $d'$  betas appeared almost unaffected by the preprocessing

598 pipeline, even though  $d'$  was computed after the outlier exclusion. This indicates that  
599 researchers who are interested in the correctness rather than the speed of responses can  
600 choose a simple preprocessing pipeline without risking false positives through elaborate  
601 transformations.

## 602 Differences between NFC groups

603 The majority of participants (61.20 %) had a preference for 1-back over the other  
604 levels, but that also means that there were 34.50 % who had a preference for 2-, 3-, or  
605 4-back, and 4.30 % who preferred no specific level over all others. It shows that when given  
606 the choice, there is a number of participants who do not prefer the easiest level, confirming  
607 the necessity of an effort discounting paradigm that works independent of the objective  
608 task load. The CAD paradigm provides the means to depict these preferences.

609 In the analysis of SV difference scores, the NFC group did not reach significance as a  
610 predictor. Conceptually, this was likely due to the partial but not full overlap between  
611 self-reported and behavioural effort investment, which has also been found for a demand  
612 selection task<sup>53</sup>. Another possible reason is the bandwidth of SVs of participants with NFC  
613 scores around the median, and the fact that the difference appeared most pronounced for  
614 4-back, and we only registered analyses of the difference scores between 1- and 2-back and  
615 2- and 3-back. As the exploratory analyses showed, a median split of NFC scores yielded a  
616 significant group difference in SVs for 4-back only, while predicting SVs with NFC as a  
617 continuous covariate showed a difference in the slopes of 1-back and 4-back. The analysis of  
618 NASA-TLX scores showed that the sum score increased with every  $n$ -back level, and that  
619 participants with NFC scores below the median had higher NASA-TLX scores for 3- and  
620 4-back than those below the median. This demonstrates that higher  $n$ -back levels have a  
621 higher discriminatory power regarding inter-individual differences in subjective effort  
622 perception. This was also supported by the fact that higher  $n$ -back levels were perceived as  
623 more aversive, and participants with NFC scores below the median reported higher

aversion than those with NFC scores above the median. Our data supports the notion of a Nonlinear Interaction between Person and Situation that has also been described by Schmitt et al. (2013)<sup>54</sup> and Blum et al. (2018)<sup>55</sup> in the same-named NIPS model. The NIPS model describes behaviour as a function of situational affordance which is mediated by personality traits. The behavioural variability follows an s-shaped curve, such that “strong” situations with low or high situational affordance elicit the least behavioural variability, while “weak” situations with moderate affordance maximize individual differences. These differences are caused by a person’s expression of a certain trait, which shifts the curve along the y-axis. In our study, the situational affordance is the *n*-back level and the behaviour is the SV, following a declining logistic curve, i.e. a mirrored s-shape. Hence, the variability in SVs increased from 1- to 4-back, and participants with higher NFC showed a more shallow decline in SVs as the situational affordance approached moderate values. According to the NIPS model, we can expect the SVs of participants with higher and lower NFC to converge again in levels of  $n > 4$ , since behavioural variability decreases when situational affordance is high. An investigation of this relationship using the COG-ED paradigm<sup>7</sup> had been encouraged by Strobel et al.<sup>53</sup> based on their findings on demand avoidance and cognitive effort investment. With the CAD paradigm, the declining logistic contrast of SVs across levels resembles the ascending logistic curve of the NIPS model<sup>54,55</sup> and should be tested further in a setting with *n*-back levels exceeding  $n = 4$ .

#### 643 Limitations

When developing a new paradigm, it is challenging to decide on the optimal analysis strategy, as every hypothesis is based on expected data patterns rather than previous findings. While the Stage 1 review process made the analyses as robust as possible, there were still unknown factors that should be addressed by future studies. For instance, the differences between participants with higher and lower NFC should be investigated with extreme groups or as a continuous variable rather than with a median split, especially in

650 academic samples where NFC can be expected to be higher on average and more narrow in  
651 range. To arrive at a sample with more balanced NFC scores, recruitment efforts should be  
652 focused on representative population samples and/or collecting data with an NFC-based  
653 stop rule. Additionally, we expected the SVs of participants with lower NFC scores to peak  
654 at 1-back and the SVs of those with higher scores to peak at 2-back, but the way the SVs of  
655 both groups appeared to drift apart in the higher  $n$ -back levels suggests that an analysis of  
656 those levels would be more fruitful in determining group differences. Future studies could  
657 create a stronger separation between the concepts investigated in this study (discounting  
658 curve, effort perception, performance, SV computation, NFC), and model the SVs and  
659 their task-related influencing factors first, before looking at (non-linear) associations with  
660 personality. Another important point is the instruction, not just for the  $n$ -back task, but  
661 for the effort discounting as well. We had to exclude several participants for  
662 misunderstanding the task instruction, so we will add a visual instruction and/or a training  
663 next time. And even though the participants were instructed to do the effort discounting  
664 with the aim to be satisfied with their choices instead of trying to increase the rewards, we  
665 cannot be sure that they did so. One might also argue that the 2€ reward range was not  
666 large enough to be an incentive for effort expenditure. However, findings by Bialaszek et  
667 al.<sup>56</sup> suggest that participants are actually more sensitive to effort when the reward is  
668 small. Nevertheless, we exceeded the largest required sample size by 2.20 times, which  
669 gives our analyses high statistical power.

## 670 Conclusion

671 Effort and reward are relevant in everyday life, yet these constructs vary in their  
672 conceptualization across individuals and even studies. With each decision an individual  
673 makes, they must weigh the required effort against the expected reward to decide if and  
674 how to behave in that situation. So far, effort discounting paradigms have relied on the  
675 assumption that the task that is objectively easiest is the one that is preferred by everyone,

and each more difficult task is simply being devalued compared to the easy one. However, effort-related traits such as Need for Cognition suggest that this is not the case. Therefore, we developed a paradigm that allows to examine effort discounting independent of objective task load, which we tested using an *n*-back task. Despite the fact that the task design allowed individuals to express a preference for higher over lower objective load levels, the overall subjective values took the shape of a declining logistic curve across *n*-back levels. The majority of participants showed a decline in subjective values at higher effort levels. A minority of participants deviated from this pattern and showed a clear preference for 2-, 3, or 4-back over 1-back. The CAD paradigm was able to depict the individual preference patterns for both those who do and do not prefer the lowest effort level. While the subjective values declined with increasing levels, they increased with better performance as measured in  $d'$ , and were unaffected by the reaction time.

Participants with Need for Cognition scores above the median reported lower subjective task load in and less aversion to more difficult levels. However, they did not have higher subjective values per se, which was due to our choice of median split and our assumption that these group differences would emerge in lower levels. The exploratory analyses showed that the predicted slope of subjective values depending on Need for Cognition scores differed between 1- and 4-back, but not between other levels. In fact, the reaction time and self-report data suggest that individual differences emerge especially from 3-back upwards, emphasizing the need for tasks with high discriminatory power and effort discounting paradigms with flexible, participant-centered mechanisms. The CAD paradigm offers this flexibility, and we encourage future studies to question traditional assumptions in the field of effort discounting in the light of these findings, and to re-use this data set for exploratory analyses.

700  
**References**

- 701 1. Botvinick, M. M., Huffstetler, S. & McGuire, J. T. Effort discounting in human  
702 nucleus accumbens. *Cognitive, affective & behavioral neuroscience* **9**, 16–27 (2009).
- 703 2. Kool, W. & Botvinick, M. Mental labour. *Nature Human Behaviour* **2**, 899–908  
704 (2018).
- 705 3. Mackworth, J. F. Paced memorizing in a continuous task. *Journal of Experimental  
706 Psychology* **58**, 206–211 (1959).
- 707 4. Jaeggi, S. M., Buschkuhl, M., Perrig, W. J. & Meier, B. The concurrent validity of  
708 the N-back task as a working memory measure. *Memory* **18**, 394–412 (2010).
- 709 5. Jonides, J. *et al.* Verbal working memory load affects regional brain activation as  
710 measured by PET. *Journal of Cognitive Neuroscience* **9**, 462–475 (1997).
- 711 6. Owen, A. M., McMillan, K. M., Laird, A. R. & Bullmore, E. N-back working memory  
712 paradigm: A meta-analysis of normative functional neuroimaging studies. *Human  
Brain Mapping* **25**, 46–59 (2005).
- 713 7. Westbrook, A., Kester, D. & Braver, T. S. What is the subjective cost of cognitive  
714 effort? Load, trait, and aging effects revealed by economic preference. *PLOS ONE*  
8, e68210 (2013).
- 715 8. Cacioppo, J. T. & Petty, R. E. The Need for Cognition. *Journal of Personality and  
716 Social Psychology* **42**, 116–131 (1982).
- 717 9. Wu, R., Ferguson, A. & Inzlicht, M. *Do humans prefer cognitive effort over doing  
718 nothing?* <https://psyarxiv.com/d2gkf/> (2021) doi:10.31234/osf.io/d2gkf.
- 719 10. Bertrams, A. & Dickhäuser, O. Passionate thinkers feel better. *Journal of Individual  
720 Differences* **33**, 69–75 (2012).

- 721 11. Nishiguchi, Y., Takano, K. & Tanno, Y. The Need for Cognition mediates and mod-  
722 erates the association between depressive symptoms and impaired Effortful Control.  
*Psychiatry Research* **241**, 8–13 (2016).
- 723 12. Xu, P. & Cheng, J. Individual differences in social distancing and mask-wearing in the  
724 pandemic of COVID-19: The role of need for cognition, self-control and risk attitude.  
*Personality and Individual Differences* **175**, 110706 (2021).
- 725 13. Kramer, A.-W., Van Duijvenvoorde, A. C. K., Krabbendam, L. & Huizenga, H. M.  
726 Individual differences in adolescents' willingness to invest cognitive effort: Relation  
to need for cognition, motivation and cognitive capacity. *Cognitive Development* **57**,  
100978 (2021).
- 727 14. Crawford, J. L., Eisenstein, S. A., Peelle, J. E. & Braver, T. S. Domain-general  
728 cognitive motivation: Evidence from economic decision-making. *Cognitive Research:  
Principles and Implications* **6**, 4 (2021).
- 729 15. Culbreth, A., Westbrook, A. & Barch, D. Negative symptoms are associated with an  
730 increased subjective cost of cognitive effort. *Journal of Abnormal Psychology* **125**,  
528–536 (2016).
- 731 16. Westbrook, A., Lamichhane, B. & Braver, T. The subjective value of cognitive effort  
732 is encoded by a domain-general valuation network. *The Journal of Neuroscience* **39**,  
3934–3947 (2019).
- 733 17. Scheffel, C., Zerna, J., Gärtner, A., Dörfel, D. & Strobel, A. Estimating individual  
734 subjective values of emotion regulation strategies. (2022).
- 735 18. Simmons, J. P., Nelson, L. D. & Simonsohn, U. A 21 word solution. (2012)  
736 doi:10.2139/ssrn.2160588.
- 737 19. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behavior Research  
738 Methods* **51**, 195–203 (2019).

- 739 20. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2020).
- 740
- 741 21. RStudio Team. *RStudio: Integrated development environment for R*. (RStudio, PBC., 2020).
- 742
- 743 22. Aust, F. & Barth, M. *papaja: Create APA manuscripts with R Markdown*. (2020).
- 744
- 745 23. Singmann, H., Bolker, B., Westfall, J., Aust, F. & Ben-Shachar, M. S. *Afex: Analysis of factorial experiments*. (2021).
- 746
- 747 24. Morey, R. D. & Rouder, J. N. *BayesFactor: Computation of Bayes factors for common designs*. (2021).
- 748
- 749 25. Greiner, B. Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association* **1**, 114–125 (2015).
- 750
- 751 26. Cacioppo, J. T., Petty, R. E. & Kao, C. F. The efficient assessment of Need for Cognition. *Journal of Personality Assessment* **48**, 306–307 (1984).
- 752
- 753 27. Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F. & Schwarz, N. Need for Cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben. *Zeitschrift für Sozialpsychologie* **25**, (1994).
- 754
- 755 28. Fleischhauer, M. *et al.* Same or different? Clarifying the relationship of need for cognition to personality and intelligence. *Personality & Social Psychology Bulletin* **36**, 82–96 (2010).
- 756
- 757 29. Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *52*, 139–183 (1988).
- 758
- 759 30. Harris, P. A. *et al.* Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics* **42**, 377–381 (2009).
- 760

- 761 31. Harris, P. A. *et al.* The REDCap consortium: Building an international community  
762 of software platform partners. *Journal of Biomedical Informatics* **95**, 103208 (2019).
- 763 32. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G\*Power 3: A flexible statistical  
764 power analysis program for the social, behavioral, and biomedical sciences. *Behavior  
Research Methods* **39**, 175–191 (2007).
- 765 33. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using  
766 G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Meth-  
ods* **41**, 1149–1160 (2009).
- 767 34. Macmillan, N. A. & Creelman, C. D. Response bias: Characteristics of detection  
768 theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin* **107**,  
401–413 (1990).
- 769 35. Whelan, R. Effective analysis of reaction time data. *The Psychological Record* **58**,  
770 475–482 (2008).
- 771 36. Berger, A. & Kiefer, M. Comparison of different response time outlier exclusion meth-  
772 ods: A simulation study. *Frontiers in Psychology* **12**, 2194 (2021).
- 773 37. Lachaud, C. M. & Renaud, O. A tutorial for analyzing human reaction times: How  
774 to filter data, manage missing values, and choose a statistical model. *Applied Psy-  
cholinguistics* **32**, 389–416 (2011).
- 775 38. Dutilh, G. *et al.* Testing theories of post-error slowing. *Attention, Perception, &  
776 Psychophysics* **74**, 454–465 (2012).
- 777 39. Houtman, F., Castellar, E. N. & Notebaert, W. Orienting to errors with and without  
778 immediate feedback. *Journal of Cognitive Psychology* **24**, 278–285 (2012).
- 779 40. Singmann, H. & Kellen, D. An introduction to mixed models for experimen-  
780 tal psychology. in *New methods in cognitive psychology* 4–31 (Routledge, 2019).  
doi:10.4324/9780429318405-2.

- 781 41. Mussel, P., Ulrich, N., Allen, J. J. B., Osinsky, R. & Hewig, J. Patterns of theta  
oscillation reflect the neural basis of individual differences in epistemic motivation.  
*Scientific Reports* **6**, (2016).
- 782 42. Enders, C. K. & Tofghi, D. Centering predictor variables in cross-sectional multilevel  
models: A new look at an old issue. *Psychological Methods* **12**, 121–138 (2007).
- 783 43. Lorah, J. Effect size measures for multilevel models: Definition, interpretation, and  
784 TIMSS example. *Large-scale Assessments in Education* **6**, (2018).
- 785 44. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nature  
Human Behaviour* **4**, 1208–1214 (2020).
- 786 45. Wetzels, R., Ravenzwaaij, D. van & Wagenmakers, E.-J. Bayesian analysis. 1–11  
(2015) doi:10.1002/9781118625392.wbecp453.
- 787 46. Cohen, J. A power primer. *Psychological Bulletin* **112**, 155–159 (1992).
- 788
- 791 47. Meule, A. Reporting and interpreting working memory performance in n-back tasks.  
*Frontiers in Psychology* **8**, (2017).
- 792
- 793 48. Kool, W., McGuire, J. T., Rosen, Z. B. & Botvinick, M. M. Decision making and the  
794 avoidance of cognitive demand. *Journal of Experimental Psychology: General* **139**,  
665–682 (2010).
- 795
- 797 49. Thomson, K. S. & Oppenheimer, D. M. The “Effort Elephant” in the room: What is  
798 effort, anyway? *Perspectives on Psychological Science* **17**, 1633–1652 (2022).
- 799 50. Klein-Flügge, M. C., Kennerley, S. W., Saraiva, A. C., Penny, W. D. & Bestmann, S.  
800 Behavioral modeling of human choices reveals dissociable effects of physical effort and  
temporal delay on reward devaluation. *PLOS Computational Biology* **11**, e1004116  
(2015).

- 801 51. Massar, S. A. A., Lim, J., Sasmita, K. & Chee, M. W. L. Sleep deprivation increases  
the costs of attentional effort: Performance, preference and pupil size. *Neuropsychologia* **123**, 169–177 (2019).
- 802 52. Fernández, L. M. & Vadillo, M. A. Flexibility in reaction time analysis: Many roads  
to a false positive? *Royal Society Open Science* **7**, 190831 (2020).
- 803 53. Strobel, A. *et al.* Dispositional cognitive effort investment and behavioral demand  
avoidance: Are they related? *PLOS ONE* **15**, e0239817 (2020).
- 804 54. Schmitt, M. *et al.* Proposal of a nonlinear interaction of person and situation (NIPS)  
model. *Frontiers in Psychology* **4**, (2013).
- 805 55. Blum, G. S., Rauthmann, J. F., Göllner, R., Lischetzke, T. & Schmitt, M. The  
nonlinear interaction of person and situation (NIPS) model: Theory and empirical  
evidence. *European Journal of Personality* **32**, 286–305 (2018).
- 806 56. Białaszek, W., Marcowski, P. & Ostaszewski, P. Physical and cognitive effort dis-  
counting across different reward magnitudes: Tests of discounting models. *PLOS  
ONE* **12**, e0182353 (2017).
- 807
- 808
- 809
- 810
- 811
- 812

813

### Acknowledgements

814 This research was partly funded by the German Research Foundation (DFG) as part  
815 of the Collaborative Research Center (CRC) 940, and partly funded by centralized funds of  
816 the Faculty of Psychology at Technische Universität Dresden. The funders have/had no  
817 role in study design, data collection and analysis, decision to publish or preparation of the  
818 manuscript. The authors would like to thank Juliana Krause and Maja Hentschel for their  
819 help with data collection, and Georgia Clay for noticing and calling our attention to the  
820 fallacy of z-transforming the hit and false alarm rates.

821

### Author Contributions

822 JZ and CS contributed equally to this work. JZ, CS, and AS conceptualized the  
823 study and acquired funding. JZ and CS developed the methodology, investigated,  
824 administered the project, and wrote the software. JZ, CS, and CK did the formal analysis.  
825 JZ visualized the results. JZ and CK prepared the original draft. All authors reviewed,  
826 edited, and approved the final version of the manuscript.

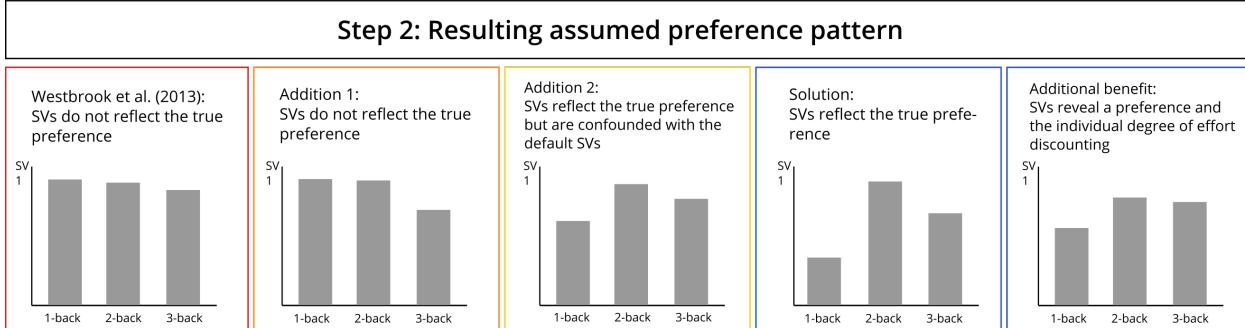
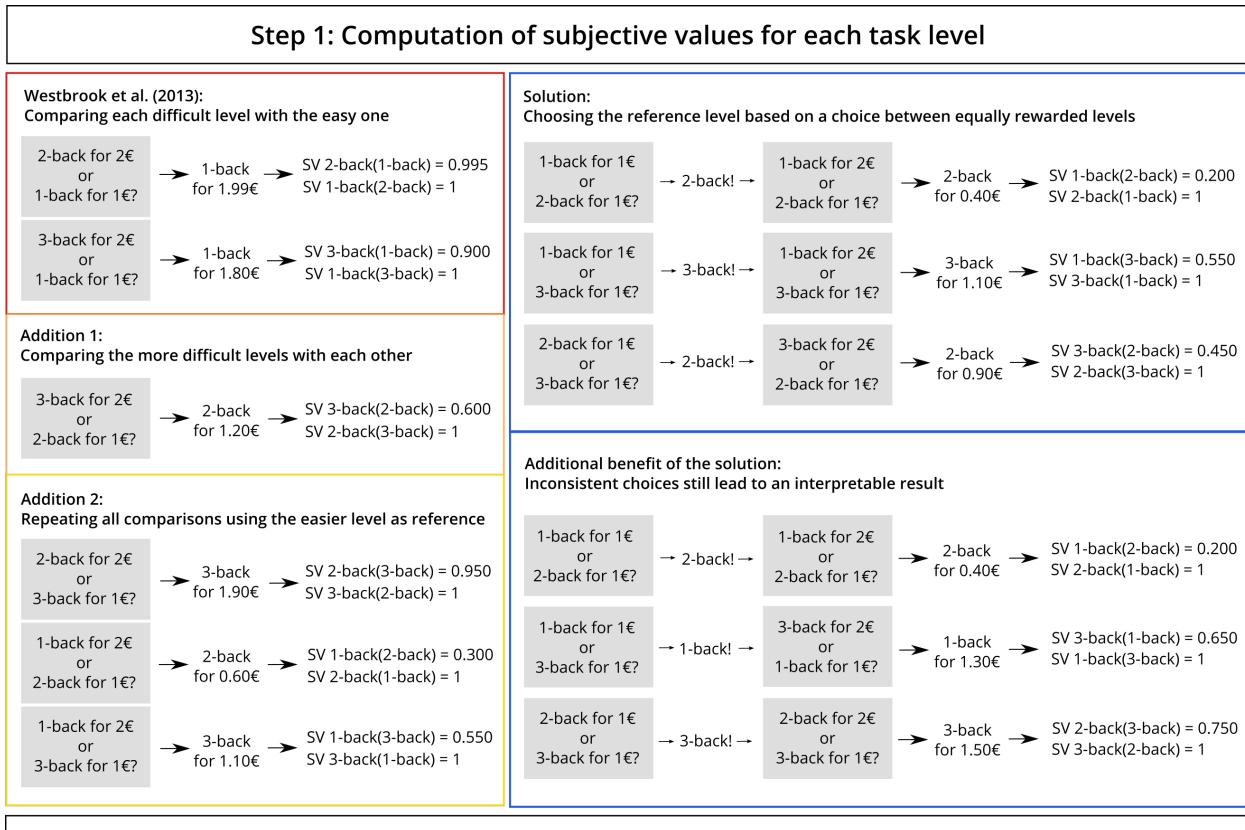
827

### Competing Interests

828

The authors declare no competing interests.

## Figures



*Figure 1.* An example for subjective values for an  $n$ -back task with three levels, returned by different modifications of the COG-ED paradigm for a hypothetical participant with the true preference  $2\text{-back} > 3\text{-back} > 1\text{-back}$ . The grey boxes are the choice options shown to the participant. The participant's final reward value of the flexible level is displayed after the first arrow. The resulting subjective value of each level is displayed after the second arrow, in the notation "SV 3-back(1-back)" for the subjective value of 3-back when 1-back is the other choice. The Solution and Additional Benefit panel follow the same logic, but are preceded by a choice between equal rewards, and the participant's first choice indicated by an exclamation mark. Figure available at [osf.io/vnj8x/](https://osf.io/vnj8x/), under a CC-BY-4.0 license.

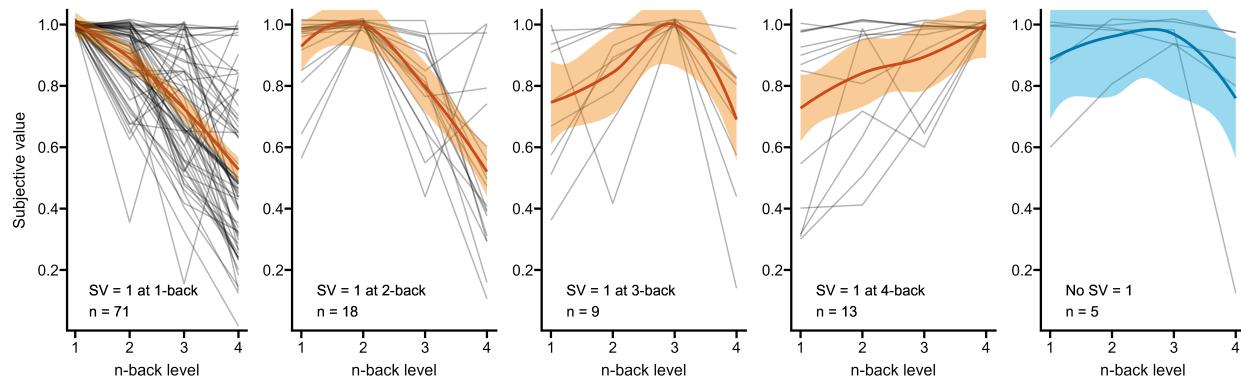
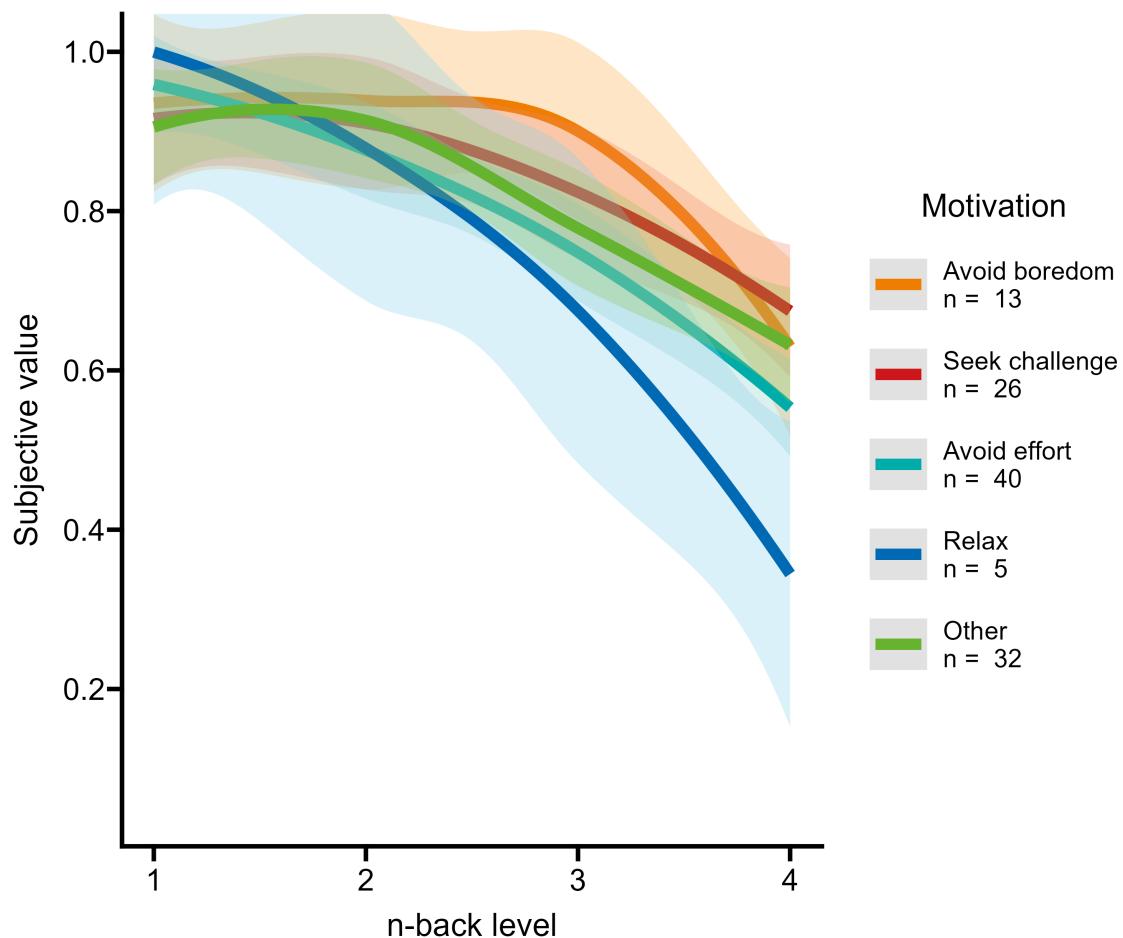
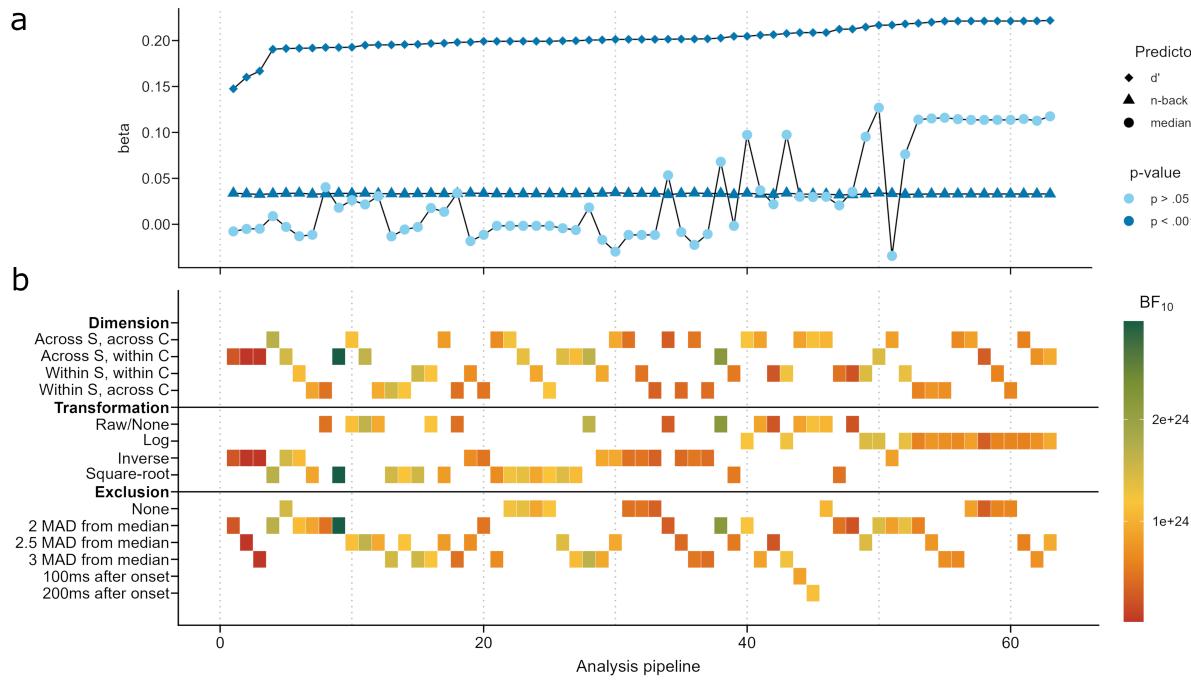


Figure 2. Subjective values (SV) per  $n$ -back level, grouped into those who had an  $SV = 1$  for 1-back, for 2-back, for 3-back, for 4-back, or no  $SV = 1$  for any level. The lines have a vertical jitter of 0.02. Smoothing of conditional means with Loess method. Transparent overlays depict the 95% confidence interval. Figure available at [osf.io/vnj8x/](https://osf.io/vnj8x/), under a CC-BY-4.0 license.



*Figure 3.* Trajectories of subjective values per *n*-back level for each participant, grouped by the motivation for effort discounting that they indicated in the single choice question after the paradigm.  $N = 116$ . 'Other' opened up a free text field. Smoothing of conditional means with Loess method. Transparent overlays depict the 95% confidence interval. Figure available at [osf.io/vnj8x/](https://osf.io/vnj8x/), under a CC-BY-4.0 license.



*Figure 4.* Results of the multi level model for each of the 63 preprocessing pipelines. Drawing a vertical through both panels indicates the type of preprocessing (panel b) of the pipeline and the resulting beta estimates of the three predictors in the model (panel a). The colourbar in panel b indicates the BF<sub>10</sub> of each multi level model compared to a model in which the n-back level has no effect. The pipelines in both panels are sorted left to right in ascending order of the magnitude of the beta estimate of the predictor  $d'$ . Figure available at [osf.io/vnj8x/](https://osf.io/vnj8x/), under a CC-BY-4.0 license.

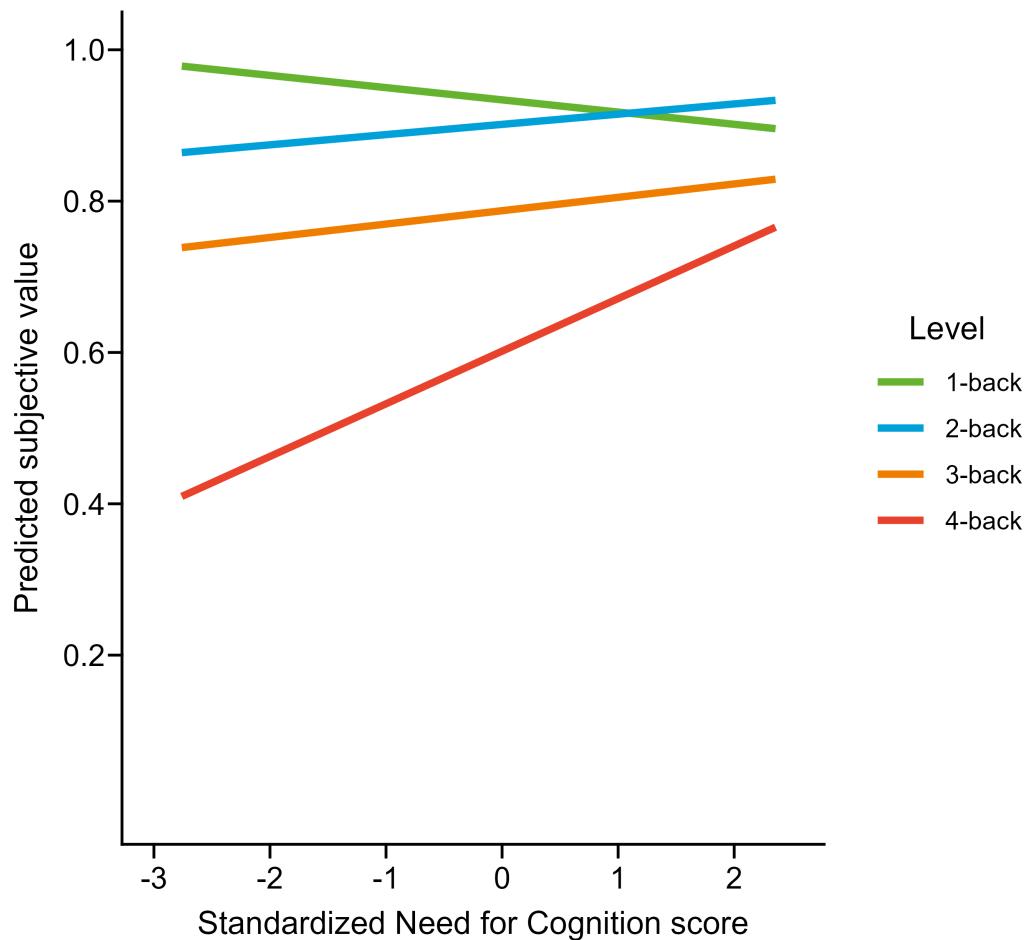


Figure 5. Predicted slopes of subjective values depending on individual Need for Cognition scores for each  $n$ -back level. The slopes of 1-back and 4-back are different at  $p = .01$ .  $N = 116$ . Figure available at [osf.io/vnj8x/](https://osf.io/vnj8x/), under a CC-BY-4.0 license.