

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

³ Josephine Zerna^{†,1}, Christoph Scheffel^{†,1}, Corinna Kührt¹, & Alexander Strobel¹

⁴ ¹ Faculty of Psychology, Technische Universität Dresden, 01062 Dresden, Germany

⁵ Author Note

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

¹⁴ Correspondence concerning this article should be addressed to Josephine Zerna,
¹⁵ 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 subjective 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 quantification of SVs for tasks without objective order of
28 task load. Results show that many participants preferred a more or the most difficult level.
29 Variance in SVs was best explained by a declining logistic contrast of the n -back levels and
30 by the accuracy of responses, while reaction time as a predictor was highly volatile
31 depending on the preprocessing pipeline. Participants with higher Need for Cognition
32 scores perceived higher n -back levels as less effortful and found them less aversive. Effects
33 of Need for Cognition on SVs in lower levels did not reach significance, as group differences
34 only emerged in higher levels. The CAD Paradigm appears to be well suited for assessing
35 and analysing task preferences independent of the supposed objective task difficulty.

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

38 Word count: 7000

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

41 **Introduction**

42 In everyday life, effort and reward are closely intertwined¹. With each decision a
43 person makes, they have to evaluate whether the effort required to reach a goal is worth
44 being exerted, given the reward they receive when reaching the goal. A reward is
45 subjectively more valuable if it is obtained with less effort, so the required effort is used as
46 a reference point for estimating the reward value¹. However, the cost of the effort itself is
47 also subjective, and research has not yet established which function best describes the
48 relationship between effort and cost². Investigating effort and cost is challenging because
49 “effort is not a property of the target task alone, but also a function of the individual’s
50 cognitive capacities, as well as the degree of effort voluntarily mobilized for the task, which
51 in turn is a function of the individual’s reward sensitivity” (p. 209)².

52 One task that is often used to investigate effort is the *n*-back task, a working memory
53 task in which a continuous stream of stimuli, e.g. letters, is presented on screen.
54 Participants indicate via button press whether the current stimulus is the same as *n* stimuli
55 before, with *n* being the level of difficulty between one and six³. The *n*-back task is well
56 suited to investigate effort because it is an almost continuous manipulation of task load as
57 has been shown by monotonic increases in error rates, reaction times⁴, and brain activity in
58 areas associated with working memory^{5,6}. However, its reliability measures are mixed, and
59 associations of *n*-back performance and measures such as executive functioning and fluid
60 intelligence are often inconsistent⁴.

61 A way to quantify the subjective cost of each *n*-back level has been developed by
62 Westbrook, Kester, and Braver⁷, called the Cognitive Effort Discounting Paradigm
63 (COG-ED). First, the participants complete the *n*-back levels to familiarize themselves
64 with the task. Then, 1-back is compared with each more difficult level by asking the

65 participants to decide between receiving a fixed 2\$ for the more difficult level or the flexible
66 starting value of 1\$ for 1-back. If they choose the more difficult level, the reward for 1-back
67 increases by 0.50\$, if they choose 1-back, it decreases by 0.50\$. This is repeated five more
68 times, with each adjustment of the 1-back reward being half of the previous step, while the
69 reward for the more difficult level remains fixed at 2\$. The idea is to estimate the point of
70 subjective equivalence, i.e., the monetary ratio at which both offers are equally preferred⁷.
71 The subjective value (SV) of each more difficult level is then calculated by dividing the
72 final reward value of 1-back by the fixed 2\$ reward. Westbrook et al.⁷ used these SVs to
73 investigate inter-individual differences in effort discounting. Younger participants showed
74 lower effort discounting, i.e., they needed a lower monetary incentive for choosing the more
75 difficult levels over 1-back.

76 The individual degree of effort discounting in the study by Westbrook et al.⁷ was also
77 associated with the participants' scores in Need for Cognition (NFC), a personality trait
78 describing an individual's tendency to actively seek out and enjoy effortful cognitive
79 activities⁸. Westbrook et al.⁷ conceptualized NFC as a trait measure of effortful task
80 engagement, providing a subjective self-report of effort discounting for each participant
81 which could then be related to the SVs as an objective measure of effort discounting. On
82 the surface, this association stands to reason, as individuals with higher NFC are more
83 motivated to mobilize cognitive effort because they perceive it as intrinsically rewarding.
84 Additionally, it has been shown that individuals avoid cognitive effort only to a certain
85 degree, possibly to retain a sense of self-control⁹, a trait more prominent in individuals
86 with high NFC^{10–12}. However, the relation of NFC and SVs might be confounded, since
87 other studies utilizing the COG-ED paradigm found the association of NFC and SVs to
88 disappear after correcting for performance¹³ or found no association of NFC and SVs at
89 all¹⁴. On the other hand, task load has been shown to be a better predictor of SVs than
90 task performance^{7,15,16}, so more research is needed to shed light on this issue.

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

92 COG-ED paradigm: That the easiest n -back level has the highest SV. We therefore
93 adapted the COG-ED paradigm in a way that allows the computation of SVs for different
94 n -back levels without presuming that all individuals inherently prefer the easiest level.
95 Since we also aim to establish this paradigm for the assessment of tasks with no objective
96 task load, e.g., emotion regulation tasks¹⁷, we call it the Cognitive and Affective
97 Discounting Paradigm (CAD). In the present study, we validated the CAD paradigm by
98 conceptually replicating the findings of Westbrook et al.⁷. Additionally, we compared the
99 effort discounting behavior of participants regarding the n -back task and an emotion
100 regulation task. The full results of the latter are published in a second Registered Report¹⁷.
101 The COG-ED paradigm has been applied to tasks in different domains before, showing
102 that SVs across task domains correlate¹⁴, but these tasks had an objective order of task
103 load, which is not the case for the choice of emotion regulation strategies or other
104 paradigms where there is no objective order of task load.

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

115 Methods

116 We report how we determined our sample size, all data exclusions (if any), all
117 manipulations, and all measures in the study^{cf. 18}. The paradigm was written and

¹¹⁸ presented using *Psychopy*¹⁹. We used *R*²⁰ with *R Studio*²¹ with the main packages *aferx*²²
¹¹⁹ and *BayesFactor*²³ for all our analyses.

¹²⁰ **Ethics information**

¹²¹ The study protocol complies with all relevant ethical regulations and was approved
¹²² by the ethics committee of the Technische Universität Dresden (reference number
¹²³ SR-EK-50012022). Prior to testing, written informed consent was obtained. Participants
¹²⁴ received 24€ in total or course credit for participation.

¹²⁵ **Design**

¹²⁶ **CAD Paradigm.** Figure 1 illustrates how different modifications of the COG-ED
¹²⁷ paradigm⁷ return SVs that do or do not reflect the true preference of a hypothetical
¹²⁸ participant, who likes 2-back most, 3-back less, and 1-back least (for reasons of clarity
¹²⁹ there are only three levels in the example). The COG-ED paradigm, which compares every
¹³⁰ more difficult level with 1-back sets the SV of 1-back to 1, regardless of the response
¹³¹ pattern. Adding a comparison of the more difficult levels with each other allows the SVs of
¹³² those two levels to be more differentiated, but leaves the SV of 1-back unchanged. Adding
¹³³ those same pairs again, but with the opposite assignment of fixed and flexible level, does
¹³⁴ approach the true preference, but has two disadvantages. First, the SVs are still quite alike
¹³⁵ across levels due to the fact that every more difficult level has only been compared with the
¹³⁶ easiest level, and second, having more task levels than just three would lead to an
¹³⁷ exponential increase in comparisons. Therefore, the solution lies in reducing the number of
¹³⁸ necessary comparisons by presenting only one effort discounting round for each possible
¹³⁹ pair of levels after determining for each pair which level should be fixed and which should
¹⁴⁰ be flexible. This is determined by presenting each possible pair of levels on screen with the
¹⁴¹ question “Would you prefer 1€ for level A or 1€ for level B?”. Participants respond by
¹⁴² clicking the respective on-screen button. Each pair is presented three times, resulting in 18

143 presented pairs, which are fully randomized in order and in the assignment of which level is
144 on the left or right of the screen. For each pair, the level that was chosen by the participant
145 at least two out of three times will be used as the level with a flexible value, which starts at
146 1€ and changes in every iteration. The other level in the pair will be set to a fixed value of
147 2€. Then, the effort discounting sensu Westbrook et al.⁷ begins, but with all possible pairs
148 and with the individually determined assignment of fixed and flexible level. The order in
149 which the pairs are presented is fully randomized, and each pair goes through all iteration
150 steps of adding/subtracting 0.50€, 0.25€, 0.13€, 0.06€, 0.03€, 0.02€ to/from the flexible
151 level's reward (each adjustment half of the previous one, rounded to two decimals) before
152 moving on to the next one. This procedure allows to compute SVs based on actual
153 individual preference instead of objective task load. For each pair, the SV of the flexible
154 level is 1, as it was preferred when faced with equal rewards, and the SV of the fixed level
155 is the final reward of the flexible level divided by 2€. Each level's "global" SV is calculated
156 as the mean of this level's SVs from all pairs in which it appeared. If the participant has a
157 clear preference for one level, this level's SV will be 1. If not, then no level's SV will be 1,
158 but each level's SV can still be interpreted as an absolute and relative value, so each
159 participant's effort discounting behaviour can still be quantified. The interpretation of SVs
160 in Westbrook et al.⁷ was "The minimum relative reward required for me to choose 1-back
161 over this level". So if the SV of 3-back was 0.6, the participant would need to be rewarded
162 with at least 60 % of what they are being offered for doing 3-back to do 1-back instead,
163 forgoing the higher reward for 3-back. In this study, the SV can be interpreted as "The
164 minimum relative reward required for me to choose any other level over this level".
165 Therefore, an SV of 1 indicates that this level is preferred over all others, while SVs lower
166 than 1 indicate that in at least one pair, a different level was preferred over this one.

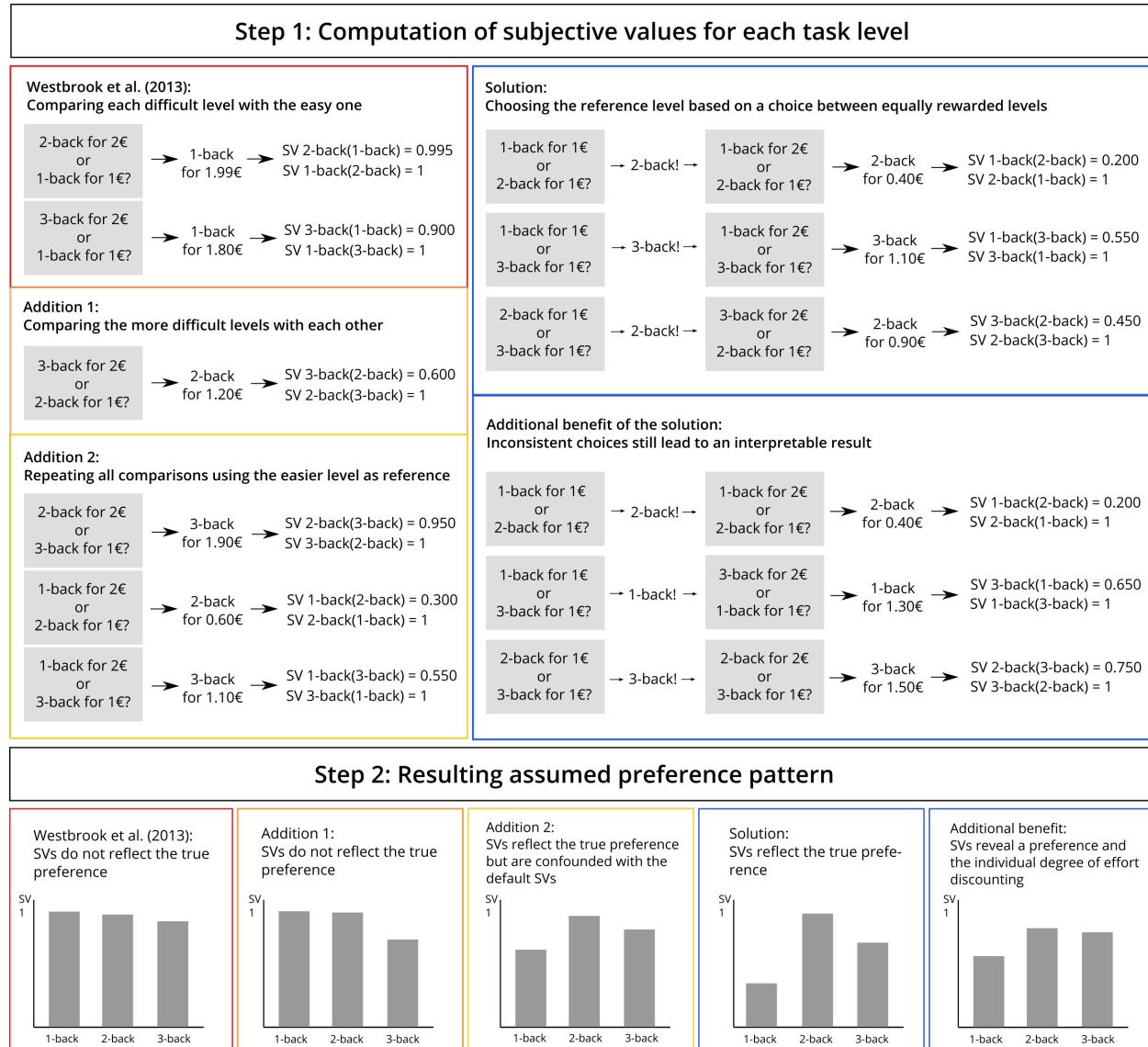


Figure 1. An example for subjective values for an n -back task with three levels, returned by different modifications of the COG-ED paradigm for a hypothetical participant with the true preference $2\text{-back} > 3\text{-back} > 1\text{-back}$. The grey boxes are the choice options shown to the participant. 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.

167 **Study procedure.** Healthy participants aged 18 to 30 years were recruited using
 168 the software *ORSEE*²⁴. Participants completed the personality questionnaires online and

then visited the lab for two sessions one week apart. NFC was assessed using the 16-item short form of the Need for Cognition Scale^{25,26}. Responses to each item (e.g., “Thinking is not my idea of fun”, recoded) were recorded on a 7-point Likert scale. The NFC scale shows comparably high internal consistency (Cronbach’s $\alpha > .80$)^{26,27}. Several other personality questionnaires were used in this study but are the topic of the Registered Report for the second lab session¹⁷. A full list of measures can be found in our Github repository. In the first session, participants provided informed consent and demographic data before completing the computer-based paradigm. The paradigm started with the *n*-back levels one to four, presented sequentially with two runs per level, consisting of 64 consonants (16 targets, 48 non-targets) per run. The levels were referred to by color (1-back: black, 2-back: red, 3-back: blue, 4-back: green) to avoid anchor effects in the effort discounting procedure. To assess perceived task load, we used the 6-item NASA Task Load Index (NASA-TLX)²⁸, where participants evaluate their subjective perception of mental load, physical load, effort, frustration, performance, and time pressure during the task on a 20-point scale. At the end of each level, participants filled out the NASA-TLX on a tablet, plus an item with the same response scale, asking them how aversive they found this *n*-back level. After the *n*-back task, participants completed the CAD paradigm on screen and were instructed to do so as realistically as possible, even though the displayed rewards were not paid out on top of their compensation. They were told that one of their choices would be randomly picked for the final run of *n*-back. However, this data was not analyzed as it only served to incentivise truthful behavior and to stay close to the design of Westbrook et al.⁷. After the CAD paradigm, participants filled out a short questionnaire on the tablet, indicating whether they adhered to the instructions (yes/no) and what the primary motivation for their decisions during the effort discounting procedure was (avoid boredom/relax/avoid effort/seek challenge/other).

The second session consisted of an emotion regulation task with negative pictures and the instruction to suppress facial reactions, detach cognitively from the picture content,

and distract oneself, respectively. The paradigm followed the same structure of task and effort discounting procedure, but participants could decide which strategy they wanted to reapply in the last block. Study data was collected and managed using REDCap electronic data capture tools hosted at Technische Universität Dresden^{29,30}.

Sampling plan

Sample size determination was mainly based on the results of the analyses of Westbrook et al.⁷ (see Design Table in the Supplementary Material). The hypothesis that yielded the largest necessary sample size was a repeated measures ANOVA with within-between interaction of NFC and *n*-back level influencing SVs. Sample size analysis with *G*Power*^{31,32} indicated that we should collect data from at least 72 participants, assuming $\alpha = .05$ and $\beta = .95$. However, the sample size analysis for the hypotheses of the second lab session revealed a larger necessary sample size of 85 participants to find an effect of $d = -0.32$ of emotion regulation on facial muscle activity with $\alpha = .05$ and $\beta = .95$. To account for technical errors, noisy physiological data, or participants who indicate that they did not follow the instructions, we aimed to collect about 50% more data sets than necessary, $N = 120$ in total.

Analysis plan

Data collection and analysis were not performed blind to the conditions of the experiments. We excluded the data of a participant from all analyses, if the participant stated that they did not follow the instructions, if the investigator noted that the participant misunderstood the instructions, or if the participant withdrew their consent. No data was replaced. The performance measure d' was computed as the difference of the *z*-transformed hit rate and the *z*-transformed false alarm rate³³. Reaction time (RT) data was trimmed by excluding all trials with responses faster than 100 ms, as the relevant cognitive processes cannot have been completed before^{34,35}. Aggregated RT values were

described using the median and the median of absolute deviation (*MAD*) as robust estimates of center and variability, respectively³⁶. Error- and post-error trials were excluded, because RT in the latter is longer due to more cautious behavior^{37,38}. To test our hypotheses, we performed a series of rmANOVAs and an MLM with orthogonal sum-to-zero contrasts in order to meaningfully interpret results³⁹.

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

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

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

247 between persons. Second, we estimated a random slopes model of n -back level (level 1,
 248 fixed, and random factor: 0-back, 1-back, 2-back, 3-back) predicting SV nested within
 249 subjects. As Mussel et al.⁴⁰ could show, participants with high versus low NFC not only
 250 have a more shallow decline in performance with higher n -back levels, but show a
 251 demand-specific increase in EEG theta oscillations, which has been associated with mental
 252 effort. We controlled for performance, i.e., d' (level 1, fixed factor, continuous), median RT
 253 (level 1, fixed factor, continuous) in order to eliminate a possible influence of declining
 254 performance on SV ratings.

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

255 Level-1-predictors were centered within cluster as recommended by Enders & Tofighi⁴¹. By
 256 this, the model yields interpretable parameter estimates. If necessary, we adjusted the
 257 optimization algorithm to improve model fit. We visually inspected the residuals of the
 258 model for evidence to perform model criticism. This was done by excluding all data points
 259 with absolute standardized residuals above 3 SD. As effect size measures, we calculated
 260 pseudo R^2 for our model and f^2 to estimate the effect of n -back level according to Lorah⁴².

261 The association of SVs and NFC was examined with an rmANOVA. We subtracted
 262 the SV of 1- from 2-back and 2- from 3-back, yielding two SV difference scores per
 263 participant. The sample was divided into participants with low and high NFC using a
 264 median split. We then computed an rmANOVA with the within-factor n -back level and the
 265 between-factor NFC group to determine whether there is a main effect of level and/or
 266 group, and/or an interaction between level and group on the SV difference scores. Post-hoc
 267 tests were computed depending on which effect reached significance at $p < .01$. To ensure
 268 the validity of this association, we conducted a specification curve analysis⁴³, which
 269 included 63 possible preprocessing pipelines of the RT data. These pipelines specify which
 270 transformation was applied (none, log, inverse, or square-root), which outliers were

271 excluded (none, 2, 2.5, or 3 *MAD* from the median, RTs below 100 or 200 ms), and across
272 which dimensions the transformations and exclusions were applied (across/within subjects
273 and across/within *n*-back levels). The rmANOVA was run with each of the 63 pipelines,
274 which also included our main pipeline (untransformed data, exclusion of RTs below
275 100 ms). The ratio of pipelines that lead to significant versus non-significant effects
276 provides an indication of how robust the effect actually is.

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

287 The results of each analysis was assessed on the basis of both *p*-value and the Bayes
288 factor BF_{10} , calculated with the *BayesFactor* package²³ using the default prior widths of
289 the functions *anovaBF*, *lmBF* and *ttestBF*. We considered a BF_{10} close to or above 3/10 as
290 moderate/strong evidence for the alternative hypothesis, and a BF_{10} close to or below
291 .33/.10 as moderate/strong evidence for the null hypothesis⁴⁴.

292 Pilot data

293 The sample of the pilot study consisted of $N = 15$ participants (53.3% female,
294 $M = 24.43$ ($SD = 3.59$) years old). One participant's data was removed because they
295 misunderstood the instruction. Due to a technical error the subjective task load data of

296 one participant was incomplete, so the hypotheses involving the NASA-TLX were analyzed
297 with $n = 14$ data sets. The results showed increases in subjective and objective task load
298 measures with higher n -back level. Importantly, SVs were lower for higher n -back levels,
299 but not different between 1- and 2-back, which shows that the easiest level is not
300 universally preferred. The MLM revealed n -back level as a reliable predictor of SV, even
301 after controlling for declining task performance (d' and median RT). NASA-TLX scores
302 were higher with higher n , and lower for the group with lower NFC scores, but NFC and
303 n -back level did not interact. All results are detailed in the Supplementary Material.

304 **Data availability**

305 The data of this study can be downloaded from osf.io/vnj8x/.

306 **Code availability**

307 The paradigm code, the R script for analysis, and the R Markdown file used to
308 compile this document are available at osf.io/vnj8x/.

309 **Protocol registration**

310 The Stage 1 Registered Report protocol has been approved and is available at
311 osf.io/qa2bg/.

312 **Results**

313 **Adjustments for Stage 2**

314 There were two necessary adjustments of the methods. First, we failed to update the
315 necessary sample size after the analyses changed with the first review round. Instead of the
316 72 subjects stated above, the largest minimum sample size was actually 53 subjects (see

³¹⁷ hypothesis 1b in the Design Table in the Supplementary Material). And secondly, we
³¹⁸ needed to re-assign the SCA to the MLM in hypothesis 2b, to which it belonged in the
³¹⁹ initial submission. Following the advice of the reviewers, we had separated the NFC
³²⁰ analysis from the MLM, and wanted to apply the SCA to the new NFC analysis
³²¹ (hypothesis 3a), as this was one of our main points of interest. However, since hypothesis
³²² 3a does not contain any RT data and the SCA is therefore only useful for the MLM in
³²³ hypothesis 2b, we applied it to the MLM.

³²⁴ **Sample**

³²⁵ Data was collected between the 16th of August 2022 and the 3rd of February 2023.
³²⁶ All of the $N = 124$ participants who filled out the online questionnaires came to the first
³²⁷ lab session. Based on the experimenters' notes, we excluded the data of seven participants
³²⁸ from analysis for misunderstanding the instruction of the n -back task, and the data of one
³²⁹ participant who reported that they confused the colours of the levels during effort
³³⁰ discounting. Our final data set therefore included $N = 116$ participants (83.60% female,
³³¹ $M = 22.40$ ($SD = 3$) years old), which is 2.2 times more than what the highest sample size
³³² calculation required.

³³³ **Manipulation checks**

³³⁴ We used rmANOVAs to investigate whether objective performance measures and
³³⁵ subjective task load measures changed across n -back levels. The performance measure d'
³³⁶ did not change across n -back levels ($F(2.85, 327.28) = 0.01, p = .999, \hat{\eta}_G^2 = .000, 90\% \text{ CI}$
³³⁷ [.000, .000], $\text{BF}_{10} = 3.31 \times 10^{-3}$), but the median RT did ($F(2.46, 283.05) = 98.67,$
³³⁸ $p < .001, \hat{\eta}_G^2 = .192, 90\% \text{ CI } [.130, .248], \text{BF}_{10} = 2.28 \times 10^{34}$), evidence was not in favour of
³³⁹ H1a but in favour of H1b. Specifically, the median RT was higher for the more difficult
³⁴⁰ level in every contrast, with two exceptions: It did not differ between 2- and 4-back, and it
³⁴¹ 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}	η_p^2	95%CI
1 - 2	-0.11	0.01	345.00	-11.76	0.000	1.75×10^{30}	0.29	[0.22, 1.00]
1 - 3	-0.16	0.01	345.00	-16.23	0.000	8.80×10^{45}	0.43	[0.37, 1.00]
1 - 4	-0.12	0.01	345.00	-12.47	0.000	4.79×10^{34}	0.31	[0.25, 1.00]
2 - 3	-0.04	0.01	345.00	-4.47	0.000	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	6.35×10^6	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.

342 All NASA-TLX subscale scores increased across n -back levels, so evidence was in
 343 favour of H1c. The effort subscale ($F(2.20, 253.06) = 203.82, p < .001, \hat{\eta}_G^2 = .316, 90\% \text{ CI}$
 344 $[.250, .375], \text{BF}_{10} = 2.47 \times 10^{34}$) increased across all levels, but the magnitude of change
 345 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
 346 4-back ($t(345) = -2.72, p_{\text{Tukey}(4)} = .035, \text{BF}_{10} = 174.38$). Three subscales had significant
 347 differences between all contrasts except for 3- versus 4-back: While ratings on the
 348 frustration and time subscales were higher for more difficult levels ($F(2.50, 287.66) = 68.06,$
 349 $p < .001, \hat{\eta}_G^2 = .172, 90\% \text{ CI } [.112, .227], \text{BF}_{10} = 5.26 \times 10^{15}$, and $F(2.21, 254.65) = 51.08,$
 350 $p < .001, \hat{\eta}_G^2 = .117, 90\% \text{ CI } [.065, .168], \text{BF}_{10} = 3.94 \times 10^9$, respectively), ratings on the
 351 performance subscale decreased with higher n ($F(2.49, 285.97) = 95.33, p < .001,$
 352 $\hat{\eta}_G^2 = .241, 90\% \text{ CI } [.176, .299], \text{BF}_{10} = 1.55 \times 10^{24}$). Ratings on the mental subscale
 353 consistently increased across all levels ($F(1.99, 228.35) = 274.47, p < .001, \hat{\eta}_G^2 = .375, 90\%$
 354 $\text{CI } [.309, .432], \text{BF}_{10} = 1.64 \times 10^{43}$). Ratings on the physical subscale were higher for more
 355 difficult levels ($F(1.68, 192.93) = 15.91, p < .001, \hat{\eta}_G^2 = .041, 90\% \text{ CI } [.009, .075],$
 356 $\text{BF}_{10} = 60.54$), apart from the contrasts 2- versus 3-back ($t(345) = -2.34, p_{\text{Tukey}(4)} = .092,$
 357 $\text{BF}_{10} = 10.45$) and 3- versus 4-back ($t(345) = -1.07, p_{\text{Tukey}(4)} = .705, \text{BF}_{10} = 0.47$). The
 358 full results of these manipulation checks are listed in Table S.1 to S.8 in the Supplementary

³⁵⁹ Material.

³⁶⁰ **Decline of subjective values**

³⁶¹ When asking participants what motivated their decisions in the effort discounting,
³⁶² 11.2% stated that they wanted to avoid boredom, 22.4% stated that they wanted a
³⁶³ challenge, 34.5% stated that they wanted to avoid effort, and 4.3% stated that they wanted
³⁶⁴ to relax. The remaining 27.6% of participants used the free text field and provided reasons
³⁶⁵ such as “I wanted a fair relation of effort and reward.”, “I wanted the fun that I had in the
³⁶⁶ more challenging levels.”, “I wanted to maximize reward first and minimize effort second.”,
³⁶⁷ or “I did not want to perform poorly when I was being paid for it.”. Figure S.1 in the
³⁶⁸ Supplementary Material shows the different motivations in the context of the SVs per
³⁶⁹ *n*-back level.

³⁷⁰ The rmANOVA showed a significant difference between the SVs across *n*-back levels
³⁷¹ ($F(1.98, 227.98) = 65.65, p < .001, \eta^2_G = .288, 90\% \text{ CI } [.222, .347], \text{BF}_{10} = 1.58 \times 10^{64}$), so
³⁷² evidence was in favour of H2a. All four pre-defined contrasts reached significance (Table 2),
³⁷³ 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	η^2_p	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.

³⁷⁴ The declining logistic contrast had the highest effect estimate ($t(345) = 12.97,$
³⁷⁵ $p < .001$), suggesting a shallow decline of SVs between 1- and 2-back, and 3- and 4-back,
³⁷⁶ respectively, and a steeper decline of SVs between 2- and 3-back.

377 Consequently, we had to adapt the MLM to incorporate this non-linear trend. To
 378 apply the contrast to the n -back levels, we had to turn the variables into a factor, with two
 379 consequences: Centered variables cannot be turned into factors, so we entered the variable
 380 level in its raw form, and factors cannot be used as random slopes, so the model is now
 381 defined as:

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

382 This means that the intercept still varied between subjects, but there were no random
 383 slopes anymore. To provide more than one observation per factor level, we used the two
 384 rounds per n -back level per subject, rather than n -back levels per subject. The ICC of the
 385 null model indicated that there was a correlation of $r = .10$ between the SVs of a subject,
 386 i.e. that 9.59% of variance in SVs could be explained by differences between participants.
 387 We did not use an optimization algorithm to improve the fit of the random intercept
 388 model. A total of 9 data points from 6 participants were excluded, because the residuals
 389 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	114.82	78.34	<.001		0.09
n-back level	0.05	0.00	799.38	18.22	<.001	0.64	
d'	0.02	0.00	798.75	5.60	<.001	0.04	
median RT	0.02	0.07	798.58	0.30	0.768	0.00	

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

390 An exploratory ANOVA was used to compare the fit of the final model with a linear
 391 random intercept model, confirming that the two models were different from each other
 392 ($\chi^2(2) = 34.48, p < .001$), and with an Akaike Information Criterion of $AIC = -492.61$
 393 and a Bayesian Information Criterion of $BIC = -454.02$ the declining logistic model was
 394 superior to the linear model ($AIC = -462.12, BIC = -433.18$). The final model had an

395 effect size of $f^2 = 0.64$ for the n -back levels and $f^2 = 0.04$ for d' , which are considered large
 396 and small, respectively⁴⁵. This means that the n -back level explained 64.20% and d'
 397 explained 3.95% of variance in SVs relative to the unexplained variance, respectively. The
 398 beta coefficient indicated that with every 1-unit increase in d' , the SV increased by 0.02.
 399 Due to the coding scheme of the logistic contrast, the beta coefficient of the n -back level
 400 has to be interpreted inversely, so SVs decline with increasing n -back level. The effect size
 401 of the median RT was $f^2 = 0.00$. Since SVs decline with increasing level, beyond the
 402 variance explained by d' , evidence was in favour of H2b.

403 To investigate the dependency of the model results on the RT preprocessing, we
 404 conducted a specification curve analysis (Figure 2).

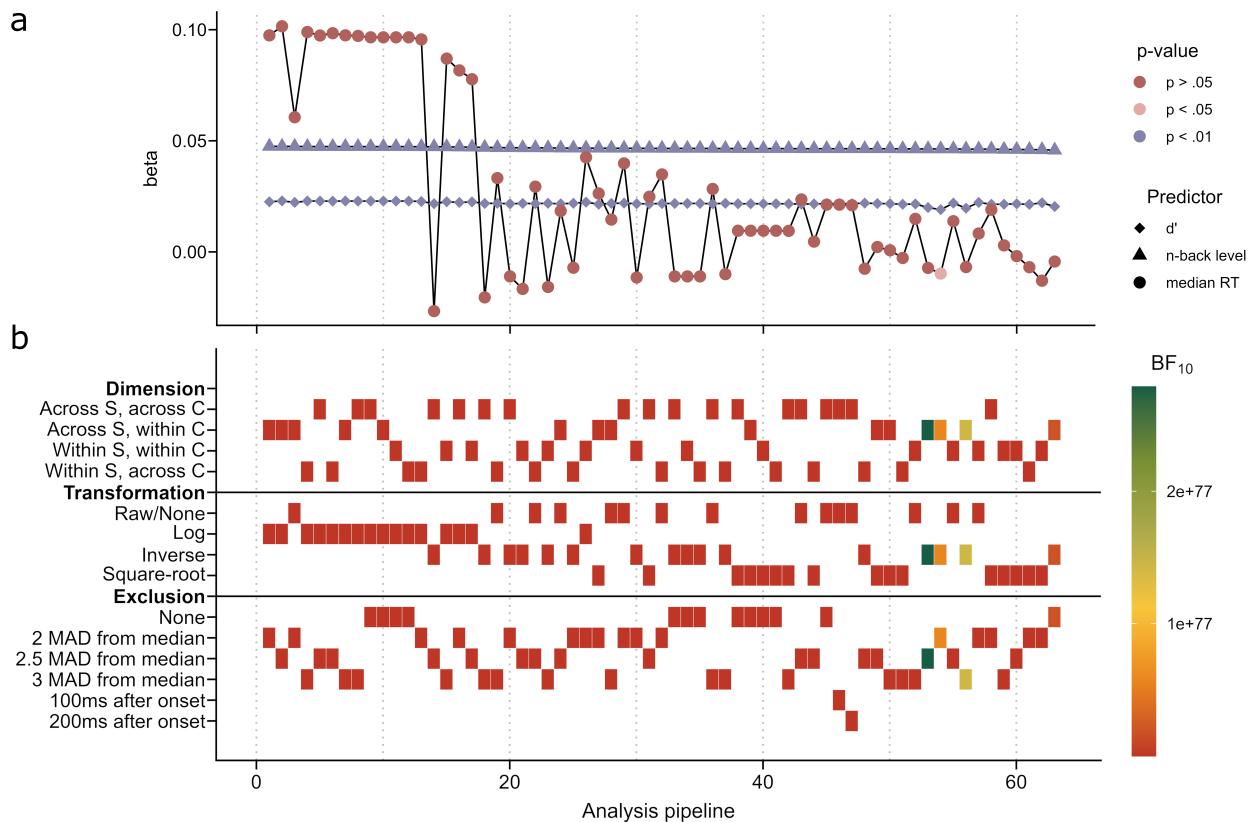


Figure 2. Results of the multi-level model for each of the 63 preprocessing pipelines. The lower panel indicates the type of preprocessing, the upper panel shows the beta coefficient of each predictor and its p-value. The colourbar indicates the BF_{10} . The pipelines are sorted in descending order of the magnitude of the n -back level beta.

405 Regardless of the preprocessing pipeline, n -back level and d' were significant
406 predictors of SVs, and had stable effect estimates across all pipelines. There was only one
407 pipeline in which the median RT was a significant predictor of SVs. This pipeline contained
408 data that had been inverse transformed across subjects but within conditions, i.e. within
409 the round of an n -back level, and RTs beyond 2 MAD from the median had been excluded.

410 **Differences between NFC groups**

411 Figure 3 shows a scatterplot of SVs per n -back level, colored depending on the
412 participant's NFC score. There is a concentration of participants who have assigned their
413 highest SV to 1-back, and this concentration fades across n -back levels. At the same time,
414 there is a subtle separation of SVs across n -back levels, depending on the participant's
415 NFC score: While the SVs of those with higher NFC scores remain elevated, the SVs of
416 those with lower NFC scores decline more strongly. Specifically, $n = 71$ participants had an
417 absolute preference for 1-back, $n = 18$ for 2-back, $n = 9$ for 3-back, and $n = 13$ for 4-back.
418 There were $n = 5$ participants who did not have an absolute preference for any n -back
419 level, i.e. none of their SVs was 1.

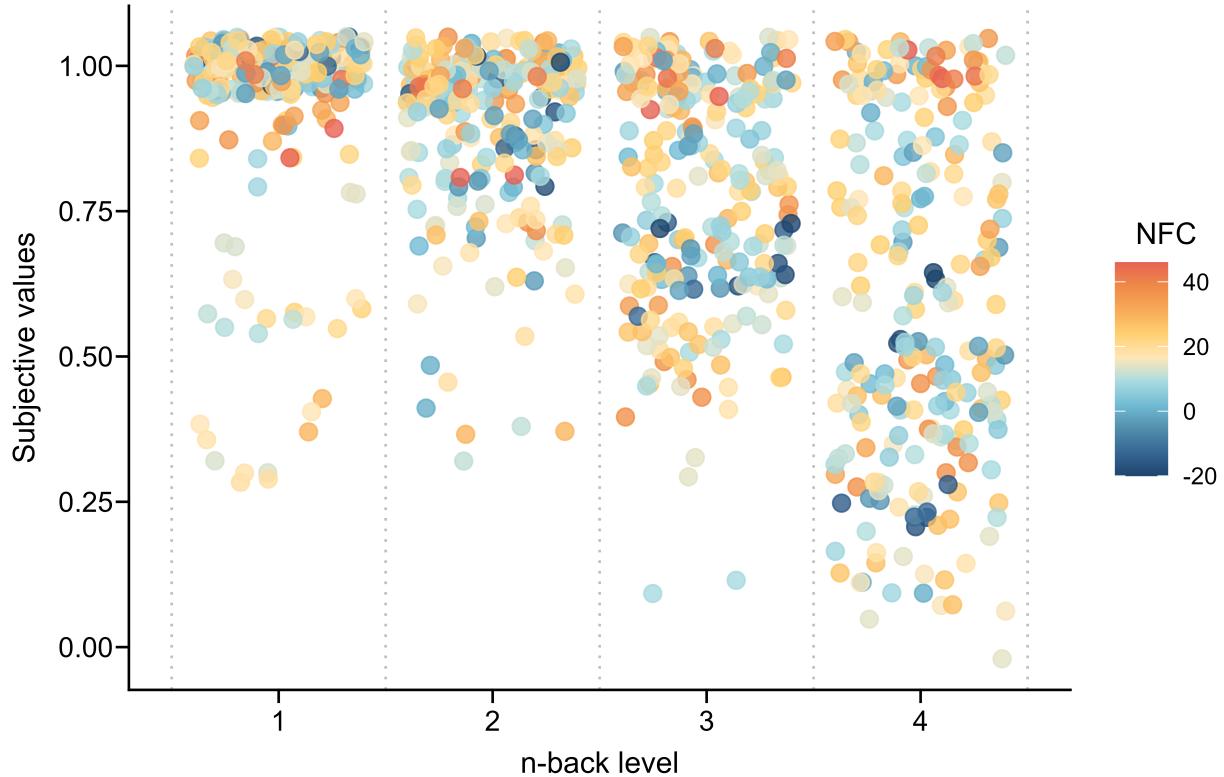


Figure 3. Subjective values per n-back level. Each dot indicates a participant, the colours indicate their Need for Cognition (NFC) score. $N = 116$. There is a horizontal jitter of 0.4 and a vertical jitter of 0.05 for visual clarity.

420 The median NFC was 16, with $n = 57$ subjects below and $n = 59$ above the median.

421 We used an rmANOVA to investigate whether the difference between the SVs of 1- and
 422 2-back, and 2- and 3-back, respectively, depended on whether a participant's NFC score
 423 was above or below the median. There was a main effect of the n -back level
 424 ($F(1, 114) = 9.13, p = .003, \hat{\eta}_G^2 = .040, 90\% \text{ CI } [.002, .115], \text{BF}_{10} = 12.68$), but neither a
 425 main effect of the NFC group ($F(1, 114) = 3.18, p = .077, \hat{\eta}_G^2 = .013, 90\% \text{ CI } [.000, .068]$,
 426 $\text{BF}_{10} = 0.56$) nor an interaction of NFC group and n -back level ($F(1, 114) = 0.46, p = .499$,
 427 $\hat{\eta}_G^2 = .002, 90\% \text{ CI } [.000, .037]$), so evidence was not in favour of H3a. Post-hoc tests
 428 showed that the difference between the SVs of 2- and 3-back is slightly more negative than
 429 the difference between 1- and 2-back ($t(114) = -3.02, p = .003$), but there were large
 430 inter-individual differences, especially for 2- and 3-back (left panel of Figure 4). This

⁴³¹ means that across the whole sample, there was a steeper decline in SVs from 2- to 3-back
⁴³² than from 1- to 2-back, but some participants showed a completely opposite pattern.

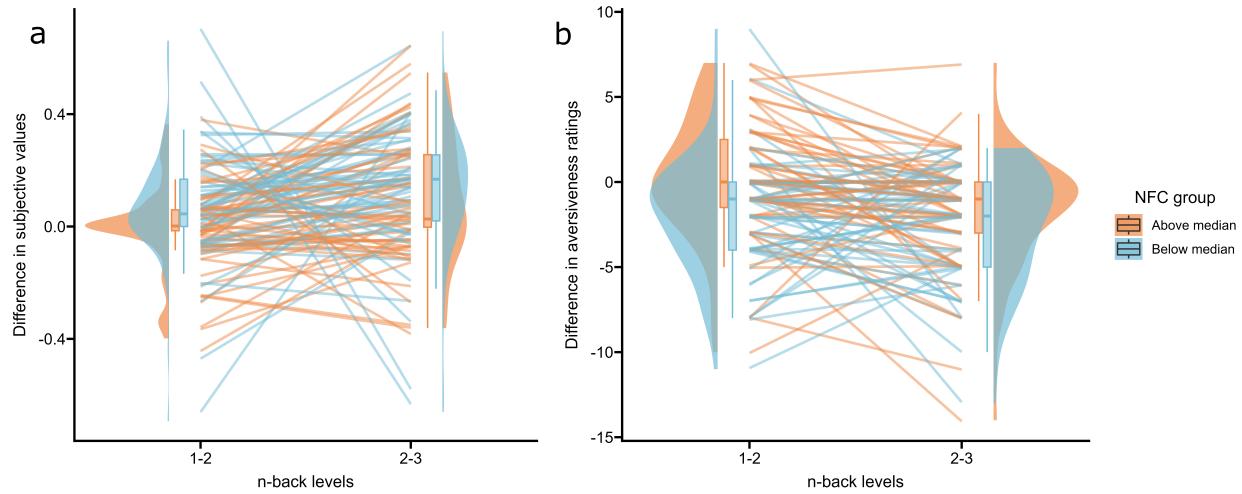


Figure 4. Difference scores for subjective values (left) and aversiveness ratings (right) when subtracting 2- from 1-back and 3- from 2-back. Each line indicates a participant, the colours indicate their Need for Cognition (NFC) score. $N = 116$.

⁴³³ The rmANOVA on the association between NFC scores and NASA-TLX scores
⁴³⁴ revealed a main effect of n -back level ($F(2.10, 239.56) = 154.50, p < .001, \hat{\eta}_G^2 = .223, 90\%$
⁴³⁵ CI [.159, .282], $BF_{10} = 2.22 \times 10^{45} \pm 0.00\%$) and an interaction between n -back level and
⁴³⁶ NFC scores ($F(2.10, 239.56) = 4.93, p = .007, \hat{\eta}_G^2 = .009, 90\% \text{ CI } [.000, .025]$), but no main
⁴³⁷ effect of NFC scores ($F(1, 114) = 3.22, p = .075, \hat{\eta}_G^2 = .022, 90\% \text{ CI } [.000, .084]$,
⁴³⁸ $BF_{10} = 1.75 \times 10^2 \pm 0.00\%$). Post-hoc tests showed that the participants with NFC scores
⁴³⁹ below the median had higher NASA-TLX scores for 3-back ($t(114) = -2.15, p = .033$,
⁴⁴⁰ $BF_{10} = 11.15$) and for 4-back ($t(114) = -2.89, p = .005, BF_{10} = 336.88$) than those with
⁴⁴¹ NFC scores above the median, so evidence was in favour of H3b. Regardless of NFC scores,
⁴⁴² NASA-TLX scores were higher for the more difficult level in each pair of n -back levels
⁴⁴³ (Figure 5).

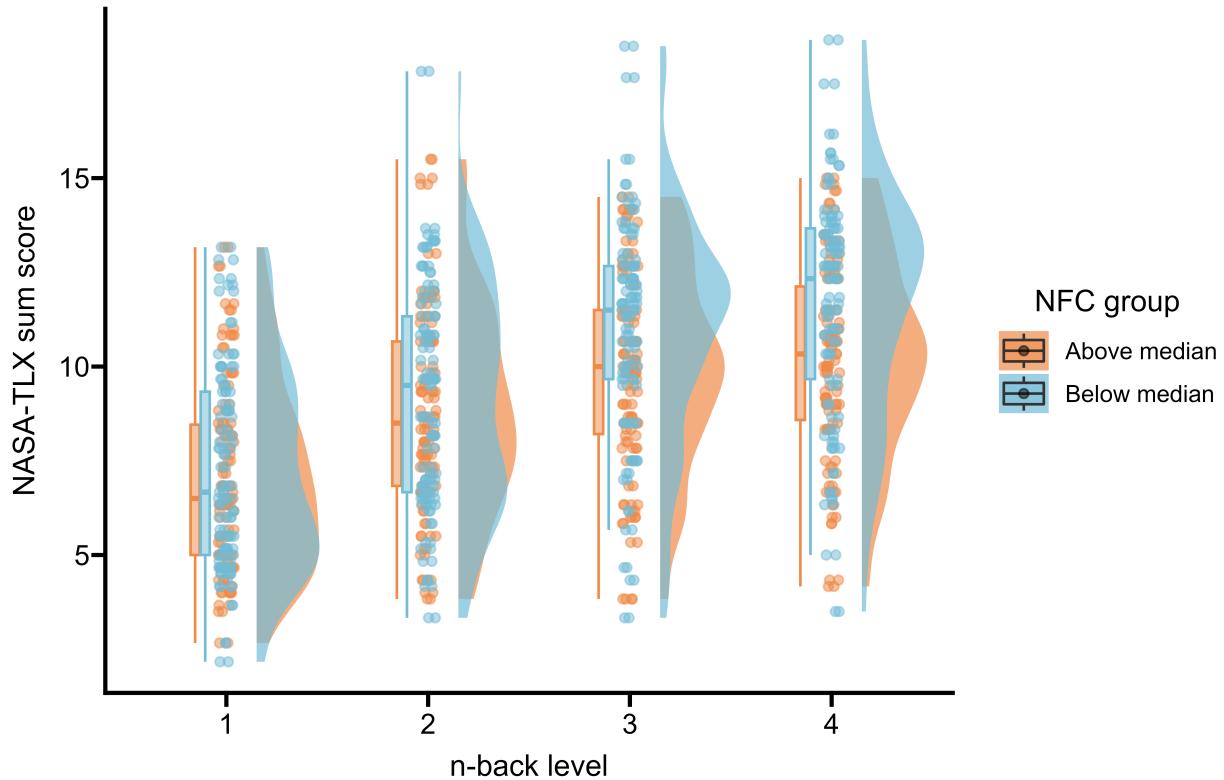


Figure 5. NASA-TLX sum scores for each n -back level. Colours indicate Need for Cognition (NFC) score above or below the median. $N = 116$.

With another rmANOVA we investigated whether the difference between the aversiveness scores of 1- and 2-back, and 2- and 3-back, respectively, depended on whether a participant's NFC score was above or below the median. There was a main effect of NFC 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 main effect of the n -back level ($F(1, 114) = 10.21, p = .002, \hat{\eta}_G^2 = .034, 90\% \text{ CI } [.000, .105]$), but no interaction ($F(1, 114) = 2.59, p = .110, \hat{\eta}_G^2 = .009, 90\% \text{ CI } [.000, .058]$). In favour of H3c, post-hoc tests revealed that participants with NFC scores below the median reported higher aversiveness than participants with NFC scores above the median ($t(114) = 2.90, p = .004$) (right panel of Figure 4). Regardless of NFC, the difference of the aversiveness scores of 2- and 3-back was smaller and more negative than that of 1- and 2-back ($t(114) = 3.20, p = .002$), but again, there were large inter-individual differences.

The full results of these analyses of NFC group differences can be found in Table S.11 to

456 S.15 in the Supplementary Material.

457 **Exploratory analysis**

458 To investigate the apparent group difference between the SVs of participants with
459 NFC scores below and above the median in higher n -back levels, we computed an
460 rmANOVA with the within-factor level (1 to 4) and the between-factor NFC group
461 (below/above median). There was no main effect of NFC group ($F(1, 114) = 2.63$,
462 $p = .108$, $\hat{\eta}_G^2 = .007$, 90% CI [.000, .053], 2.95×10^{-1}), but a main effect of the n -back level
463 ($F(2.01, 229.39) = 67.39$, $p < .001$, $\hat{\eta}_G^2 = .295$, 90% CI [.228, .354], 2.70×10^{30}) and an
464 interaction ($F(2.01, 229.39) = 3.24$, $p = .041$, $\hat{\eta}_G^2 = .020$, 90% CI [.000, .044]). Post-hoc
465 tests for the main effect of level showed that SVs were lower for the more difficult n -back
466 level in each paired contrast except for 1- versus 2-back. Post-hoc tests for the interaction
467 effect showed that the NFC groups only had a significant difference in SVs for 4-back,
468 where participants below the NFC median had lower scores ($\Delta M = 0.11$, 95% CI
469 [0.01, 0.22], $t(114) = 2.13$, $p = .036$). Despite not reaching significance, 1-back was the only
470 level in which participants with NFC scores above the median seemed to have lower SVs
471 than those with scores below the median ($\Delta M = -0.05$, 95% CI [-0.11, 0.01],
472 $t(114) = -1.50$, $p = .136$). The full results of this exploratory analysis of NFC group
473 differences can be found in Table S.16 and S.17 in the Supplementary Material.

474 **Discussion**

475 This Registered Report aimed to adapt the Cognitive Effort Discounting (COG-ED)
476 paradigm by Westbrook et al.⁷, which estimates subjective values of different n -back levels,
477 into the Cognitive and Affective Discounting (CAD) paradigm, which estimates SVs of
478 tasks without defaulting to the objective task load as a benchmark. For this purpose, we
479 changed the way in which the discounting options are presented to the participants, based
480 the anchor on their own choices, and computed SVs across multiple combinations of task

481 levels. The analyses were closely aligned with those in Westbrook et al.⁷ to demonstrate
482 the changes in SVs brought about by the new paradigm. This study also applied the CAD
483 paradigm to an emotion regulation task, the results of which are detailed in a second
484 Registered Report¹⁷.

485 **Manipulation checks**

486 The performance measure d' did not differ across n -back levels, but the RT increased
487 from 1- to 2- to 3-back and then remained on a high level for 4-back. This points to three
488 important characteristics of the n -back task in this context. Firstly, RT as a valid
489 group-level indicator of performance might only be useful for levels up to $n = 3$, and could
490 be used to investigate inter-individual differences for $n > 3$. Secondly, there is a
491 speed-accuracy tradeoff in the first three levels, that might even re-emerge in higher levels,
492 where d' would decline and RT would remain stable. And lastly, the fact that neither
493 accuracy nor speed is an informative performance measure by itself has been observed
494 before⁴⁶ and both show different associations with various measures of intelligence⁴,
495 suggesting that they should always be reported as separate indices. Additionally, d' might
496 not have differed across n -back levels because the manipulation of task load is not strictly
497 continuous. Several participants said that they perceived 3-back as more difficult than
498 4-back because they found it is easier to remember chunks of stimuli when n was an even
499 number than when n was an odd number.

500 All NASA-TLX subscales differed across n -back levels, but the effort and mental load
501 subscales were the only ones to consistently increase across all levels. This would support
502 the notion of the n -back task offering a continuous manipulation of task load, at least
503 subjectively. Ratings on the frustration and time subscales increased and ratings on the
504 performance subscale decreased until 3-back and then remained stable. This pattern is
505 akin to the RT, which also increased and then remained stable. Ratings on the physical
506 load subscale increased with n -back levels, but not between 2- and 3-back and 3- and

507 4-back, respectively.

508 **Decline of subjective values**

509 The rmANOVA with different pre-defined contrasts showed that all fit the SVs to a
510 different degree, and that the SVs do not simply decline linearly across n -back levels. The
511 best fit was a declining logistic curve, reflecting that the majority of participants preferred
512 1-back and that SVs for 2-back were also high, before having more inter-individual variance
513 for 3- and 4-back. Thomson and Oppenheimer⁴⁷ argue that the different effort curves that
514 have been observed for different tasks are likely due to the fact that we still understand
515 quite little about how and why different manipulations of effort work. For example, the
516 n -back task is likely not a continuous manipulation of task load, as discussed above.

517 However, the declining logistic curve is similar to the sigmoidal curve that had the best fit
518 in a different effort paradigm⁴⁸, which the authors explained with the low effect of low
519 energy costs, suggesting there are still common features of effort across different tasks and
520 domains. The MLM with the logistic contrast showed that the n -back level explained the
521 majority of variance in SVs, while the performance measure d' also explained some variance
522 in SVs, albeit less. With increasing n -back level and decreasing d' , the SV decreased. The
523 median RT was not a significant predictor in this model, which was somewhat surprising
524 because RT but not d' yielded significant differences across levels in the manipulation
525 checks. However, participants might have deliberately or subconsciously used the feedback
526 they received at the end of each round, i.e. twice per n -back level, as an anchor during the
527 effort discounting. This feedback was based on correct responses and not on RT, so if
528 participants based their effort discounting choices at least partly on this feedback, they
529 were either motivated to repeat a task in which they performed well and/or they were
530 reluctant to accept a larger reward for a task in which they did not perform well. Since
531 more participants reported effort avoidance as their motivation in the effort discounting
532 than those who reported seeking a challenge, we can assume that they were more

533 motivated to repeat a task in which they performed well because their good performance
534 coincided with low effort.

535 The declining logistic n -back levels and d' remained significant predictors of SVs
536 throughout all 63 preprocessing pipelines in the specification curve analysis, with betas
537 that varied by less than 0.01. In contrast to this stood the variability of the median RT
538 betas, which ranged from about 0.10 to -0.03, and reached significance in only one pipeline.
539 This pipeline was among the three pipelines with the highest BF_{10} , and applied inverse
540 transformation to the RT data, across subjects but within conditions, and excluded data
541 beyond 2 MAD from the median. Interestingly, the curve of median RT betas in the upper
542 panel of Figure 2 mirrored the rectangular pipeline indicators in the transformation rows of
543 the lower panel, so the transformation choice influenced the median RT much more than
544 the dimension or the exclusion choice did. As Fernandez et al.⁴⁹ found, applying more than
545 one preprocessing step to the reaction time data of a Stroop task increased the risk of false
546 positives beyond $\alpha = .05$, and transformation choices inflated this risk more than outlier
547 exclusion or aggregation choices did. Our data seems to corroborate this finding for n -back
548 tasks as well. Surprisingly, the d' betas appear almost unaffected by the preprocessing
549 pipeline, even though d' was computed after the outlier exclusion. This indicates that
550 researchers who are interested in the correctness rather than the speed of responses can
551 choose a simple preprocessing pipeline without risking false positives through elaborate
552 transformations.

553 Differences between NFC groups

554 The majority of participants (61.20 %) had an absolute preference for 1-back over the
555 other levels, but that also means that there were 34.50 % who had an absolute preference
556 for 2-, 3-, or 4-back, and 4.30 % who preferred no specific level over all others. It shows that
557 when given the choice, there is a large number of participants who do not prefer the easiest
558 level, confirming the necessity of an effort discounting paradigm that works independent of

559 the objective task load. The CAD paradigm provides the means to depict these preferences.

560 Despite the visual separation of the SVs of participants with very low and high NFC
561 scores in higher n -back levels, the NFC group did not reach significance in predicting SVs.
562 This was likely due to the bandwidth of SVs of participants with NFC scores around the
563 median, and due to the fact that the difference appeared most pronounced for 4-back, and
564 we only analyzed the difference scores between 1- and 2-back and 2- and 3-back. As the
565 exploratory analysis showed, only 4-back yielded a significant group difference, and SVs of
566 participants with NFC scores above the median were higher for 2- to 4-back and lower for
567 1-back. The analysis of NASA-TLX scores showed that the sum score increased with every
568 n -back level, and that participants with NFC scores below the median had higher
569 NASA-TLX scores for 3- and 4-back than those below the median. This demonstrates that
570 higher n -back levels have a higher discriminatory power regarding inter-individual
571 differences in subjective effort perception. This was also supported by the fact that higher
572 n -back levels were perceived as more aversive, and participants with NFC scores below the
573 median reported higher aversion than those with NFC scores above the median.

574 **Limitations**

575 When developing a new paradigm, it is challenging to decide on the optimal analysis
576 strategy, as every hypothesis is based on expected data patterns rather than previous
577 findings. While the Stage 1 review process made the analyses as robust as possible, there
578 were still unknown factors that should be addressed by future studies. For instance, the
579 differences between participants with higher and lower NFC should be investigated with
580 extreme groups rather than a median split, especially in academic samples where NFC can
581 be expected to be higher on average and more narrow in range. Additionally, we expected
582 the SVs of participants with lower NFC scores to peak at 1-back and the SVs of those with
583 higher scores to peak at 2-back, but the way the SVs of both groups appeared to drift
584 apart in the higher n -back levels suggests that an analysis of those levels would be more

585 fruitful in determining group differences. Future studies could create a stronger separation
586 between the concepts investigated in this study (discounting curve, effort perception,
587 performance, SV computation, NFC), and model the SVs and their task-related influencing
588 factors first, before looking at (non-linear) associations with personality. Another
589 important point is the instruction, not just for the *n*-back task, but for the effort
590 discounting as well. We had to exclude several participants for misunderstanding the task
591 instruction, so we will add a visual instruction or a training next time. And even though
592 the participants were instructed to do the effort discounting with the aim to be satisfied
593 with their choices instead of trying to increase the rewards, we cannot be sure that they
594 did so. One might also argue that the 2€ reward range was not large enough to be an
595 incentive for effort expenditure. However, findings by Bialaszek et al.⁵⁰ suggest that
596 participants are actually more sensitive to effort when the reward is small. Nevertheless, we
597 exceeded the largest required sample size by 2.20 times, which gives our analyses high
598 statistical power, and we adhered to the agreed upon analyses of Stage 1.

599 Conclusion

600 Effort and reward are two highly subjective concepts with relevance in daily life.
601 With each decision an individual makes, they must weigh the required effort against the
602 expected reward to decide if and how to behave in that situation. So far, effort discounting
603 paradigms have relied on the assumption that the task that is objectively easiest is the one
604 that is preferred by everyone, and each more difficult task is simply being devalued
605 compared to the easy one. However, effort-related traits such as Need for Cognition suggest
606 that this is not the case. Therefore, we developed a paradigm that allows effort discounting
607 independent of objective task load, which we tested using an *n*-back task. The results
608 showed that many participants indeed preferred a more or even the most difficult *n*-back
609 level. Spanning the entire sample, these preferences took the shape of a declining logistic
610 curve across *n*-back levels. While the subjective value declined with increasing levels, it

increased with better performance as measured in d' , and was 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 likely due to our choice of median split and our assumption that these group differences would emerge in lower 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.

622

References

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

642

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

- 661 20. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2020).
- 662
- 663 21. RStudio Team. RStudio: Integrated development for R. (2020).
- 664
- 665 22. Singmann, H., Bolker, B., Westfall, J., Aust, F. & Ben-Shachar, M. S. *Afex: Analysis of factorial experiments*. (2021).
- 666
- 667 23. Morey, R. D. & Rouder, J. N. *BayesFactor: Computation of Bayes factors for common designs*. (2021).
- 668
- 669 24. Greiner, B. Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association* **1**, 114–125 (2015).
- 670
- 671 25. Cacioppo, J. T., Petty, R. E. & Kao, C. F. The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment* **48**, 306–307 (1984).
- 672
- 673 26. 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).
- 674
- 675 27. 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).
- 676
- 677 28. Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *52*, 139–183 (1988).
- 678
- 679 29. 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).
- 680
- 681 30. Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics* **95**, 103208 (2019).
- 682

- 683 31. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*Power 3: A flexible statistical
power analysis program for the social, behavioral, and biomedical sciences. *Behavior
Research Methods* **39**, 175–191 (2007).
- 684 32. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using
G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Meth-
ods* **41**, 1149–1160 (2009).
- 685 33. Macmillan, N. A. & Creelman, C. D. Response bias: Characteristics of detection
theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin* **107**,
401–413 (1990).
- 686 34. Whelan, R. Effective Analysis of Reaction Time Data. *The Psychological Record* **58**,
475–482 (2008).
- 687 35. Berger, A. & Kiefer, M. Comparison of Different Response Time Outlier Exclusion
Methods: A Simulation Study. *Frontiers in Psychology* **12**, 2194 (2021).
- 688 36. Lachaud, C. M. & Renaud, O. A tutorial for analyzing human reaction times: How
to filter data, manage missing values, and choose a statistical model. *Applied Psy-
cholinguistics* **32**, 389–416 (2011).
- 689 37. Dutilh, G. *et al.* Testing theories of post-error slowing. *Attention, Perception, &
Psychophysics* **74**, 454–465 (2012).
- 690 38. Houtman, F., Castellar, E. N. & Notebaert, W. Orienting to errors with and without
immediate feedback. *Journal of Cognitive Psychology* **24**, 278–285 (2012).
- 691 39. Singmann, H. & Kellen, D. An introduction to mixed models for experimen-
tal psychology. in *New methods in cognitive psychology* 4–31 (Routledge, 2019).
doi:10.4324/9780429318405-2.
- 692 40. 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).

- 702
- 703 41. Enders, C. K. & Tofghi, D. Centering predictor variables in cross-sectional multilevel
704 models: A new look at an old issue. *Psychological Methods* **12**, 121–138 (2007).
- 705 42. Lorah, J. Effect size measures for multilevel models: Definition, interpretation, and
706 TIMSS example. *Large-scale Assessments in Education* **6**, (2018).
- 707 43. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nature
708 Human Behaviour* **4**, 1208–1214 (2020).
- 709 44. Wetzels, R., Ravenzwaaij, D. van & Wagenmakers, E.-J. Bayesian analysis. 1–11
710 (2015) doi:10.1002/9781118625392.wbcp453.
- 711 45. Cohen, J. A power primer. *Psychological Bulletin* **112**, 155–159 (1992).
- 712
- 713 46. Meule, A. Reporting and interpreting working memory performance in n-back tasks.
714 *Frontiers in Psychology* **8**, (2017).
- 715 47. Thomson, K. S. & Oppenheimer, D. M. The “Effort Elephant” in the room: What is
716 effort, anyway? *Perspectives on Psychological Science* **17**, 1633–1652 (2022).
- 717 48. Klein-Flügge, M. C., Kennerley, S. W., Saraiva, A. C., Penny, W. D. & Bestmann, S.
718 Behavioral modeling of human choices reveals dissociable effects of physical effort and
temporal delay on reward devaluation. *PLOS Computational Biology* **11**, e1004116
(2015).
- 719 49. Fernández, L. M. & Vadillo, M. A. Flexibility in reaction time analysis: Many roads
720 to a false positive? *Royal Society Open Science* **7**, 190831 (2020).
- 721 50. Białaszek, W., Marcowski, P. & Ostaszewski, P. Physical and cognitive effort dis-
722 counting across different reward magnitudes: Tests of discounting models. *PLOS
ONE* **12**, e0182353 (2017).

723

Acknowledgements

724 This research is partly funded by the German Research Foundation (DFG) as part of
725 the Collaborative Research Center (CRC) 940, and partly funded by centralized funds of
726 the Faculty of Psychology at Technische Universität Dresden. The funders have/had no
727 role in study design, data collection and analysis, decision to publish or preparation of the
728 manuscript. The authors would like to thank Juliana Krause and Maja Hentschel for their
729 help with data collection.

730

Author Contributions

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

736

Competing Interests

737

The authors declare no competing interests.