

When easy is not preferred: An effort discounting paradigm for estimating subjective
values of tasks

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

¹ Faculty of Psychology, Technische Universität Dresden, 01069 Dresden, Germany

Author Note

The authors made the following contributions. Josephine Zerna: Conceptualization, Methodology, Funding acquisition, Formal analysis, Investigation, Project administration, Software, Visualization, Writing - original draft preparation, Writing - review & editing; Christoph Scheffel: Conceptualization, Methodology, Funding acquisition, Investigation, Project administration, Software, Writing - review & editing; Corinna Kührt: Formal analysis, Writing - original draft preparation, Writing - review & editing, Visualization. [†] Josephine Zerna and Christoph Scheffel contributed equally to this work.

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

Abstract

15

16 When individuals set goals, they consider the subjective value (SV) of both the
17 anticipated reward and the required effort, a trade-off that is of great interest to
18 psychological research. However, the SV of effort is highly individual, and previous
19 quantification approaches have had two crucial limitations: They presumed a unidirectional
20 relationship between objective task load and the SV of effort, and as a consequence, they
21 could only quantify the SVs of tasks with an objective order of task load. One of these
22 approaches is the Cognitive Effort Discounting paradigm by Westbrook et al. (2013). We
23 aim to replicate their analysis with our adaptation, the Cognitive and Emotion Regulation
24 Effort Discounting (CERED) paradigm. We argue that the CERED paradigm allows two
25 crucial things: Quantifying SVs without assuming that the easiest level is preferred, and
26 quantifying SVs for tasks with no objective order of task load.

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

29 Word count: X

When easy is not preferred: An effort discounting paradigm for estimating subjective values of tasks

Introduction

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

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

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

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

The individual degree of ED in the study by Westbrook et al.⁷ was also associated with the participants' Need for Cognition (NFC) score, a personality trait describing individuals who actively seek and enjoy effortful cognitive activities.⁸ Westbrook et al.⁷ conceptualized NFC as a trait measure of effortful task engagement, providing a subjective self-report of ED for each participant which could then be related to the SVs as an objective measure of ED. On the surface, this association stands to reason, as individuals with higher NFC are more motivated to mobilize cognitive effort because they perceive it as intrinsically rewarding. Additionally, it has been shown that individuals avoid cognitive effort only to a certain degree, possibly to retain a sense of self-control,⁹ a trait more prominent in individuals with high NFC.¹⁰⁻¹² However, the relation of NFC and SVs might be confounded, since other studies utilizing the COG-ED paradigm found the association of NFC and SVs to disappear after correcting for performance¹³ or found no association of NFC and SVs at all.¹⁴ On the other hand, task load has been shown to be a better predictor of SVs than task performance,^{7,15,16} so more research is needed to shed light on this issue.

The present study changes one fundamental assumption of the original COG-ED paradigm: That the easiest n-back level has the highest SV. We adapted the COG-ED paradigm in such a way that it allows the computation of SVs for different n-back levels

without presuming that all individuals inherently prefer the easiest level. Figure 1 illustrates how different modifications of the COG-ED paradigm return SVs that do or do not reflect the true preference of a hypothetical participant, who likes 2-back most, 3-back less, and 1-back least. The COG-ED paradigm sets the SV of 1-back to 1, regardless of the response pattern. Adding a comparison of 2-back and 3-back allows the SVs of those two levels to be more differentiated, but leaves the SV of 1-back unchanged. Adding three more comparisons of the same levels but using the easier level as reference does approach the true preference, but has two disadvantages. First, the SVs are still distorted by the SVs returned by the original paradigm, and second, having more task levels would lead to an exponential increase in comparisons. Therefore, the solution lies in reducing the number of necessary comparisons by presenting only one ED round for each possible pair of levels, and by starting each round with a choice between equal prices. For example, the participant is presented with the choice of receiving 1€ for 2-back or 1€ for 4-back. The level chosen by the participant will then be used as the level with a flexible value, which starts at 1€ and is changed in every iteration. The level that was not chosen will be set to a fixed value of 2€. This procedure allows to compute SVs based on actual individual preference instead of objective task load. Each level's SV is calculated as the mean of this level's SVs from all comparisons 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 ED behaviour can still be quantified. Since we also aim to establish this paradigm for the assessment of tasks with no objective task load, e.g. emotion regulation tasks, we call it the Cognitive and Emotion Regulation Effort Discounting Paradigm (CERED). In the present study, we will validate the CERED paradigm by conceptually replicating the findings of Westbrook et al..⁷ Additionally, we will compare the ED behaviour of participants regarding the n-back task and an emotion regulation task. The full results of the latter will be published in a second Registered Report. The COG-ED paradigm has been applied to tasks with different

domains before, showing that SVs across task domains correlate,¹⁴ but these tasks had an
objective order of task load, which is not the case for emotion regulation.



Figure 1

Fig.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 participant with the true
preference 2-back > 3-back > 1-back.

Our hypotheses were derived from the results of Westbrook et al..⁷ Regarding the
associations of subjective and objective task load we hypothesize that (1a) the signal
detection parameter d' declines with increasing n-back level, (1b) reaction time increases
with increasing n-back level, and (1c) perceived task load increases with increasing n-back

level. Regarding the associations of task load and ED we hypothesize that (2a) SVs decline with increasing n-back level, and (2b) they do so even after controlling for declining task performance. Here we added the hypothesis that (2c) SVs decline stronger with increasing task load for individuals with low compared to high NFC scores. And regarding individual differences in ED we hypothesize that (3a) SVs predict individual NFC scores, and (3b) perceived task load does not predict individual NFC scores.

Methods

The paradigm was written and presented using *Psychopy*.¹⁷ We used *R Studio*^{18,19} with the main packages *afex*²⁰ and *bayestestR*²¹ for all our analyses.

Ethics information

Pilot data

The sample of the pilot study consisted of $N = 15$ participants (53.30% female, $M = 24.40$ ($SD = 3.60$) years old). One participant's data was removed because they misunderstood the instruction. Due to a technical error the subjective task load data of one participant was incomplete, so the hypotheses involving the NASA Task Load Index were analyzed with $n = 14$ data sets. The results showed increases in subjective and objective task load measures with higher n-back level. Importantly, SVs were lower for higher n-back levels, but not different between 1- and 2-back, which can be considered preliminary proof-of-concept, as this phenomenon can only emerge in this version of the paradigm. The MLM revealed that n-back level was a reliable predictor of SV, even after controlling for declining task performance (d' and RT) as well as correct and post-correct answers, while NFC was not. The specification curve analysis showed that this pattern was true for all 63 pipelines. And finally, while the $Ax\text{AUC}$ value did not predict any amount

of variance in individual NFC scores, the *AUC* of NASA-TLX scores did. All results are detailed in the Supplementary Material.

Design

Healthy participants aged 18 to 30 years will be recruited using the software *ORSEE*.²² Participants will fill out the personality questionnaires online and then visit the lab for two sessions one week apart. NFC will be assessed using the 16-item short form of the Need for Cognition Scale.^{23,24} Responses to each item (e.g., “Thinking is not my idea of fun”, recoded) will be recorded on a 7-point Likert scale. The NFC scale shows comparably high internal consistency (Cronbach’s $\alpha > .80$).^{23,25} Several other personality questionnaires will be used in this study but are the topic of the Registered Report for the second lab session. A full list of measures can be found in our Github repository. In the first session, participants provide informed consent and demographic data before completing the computer-based paradigm. The paradigm starts 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 are referred to by color (1-back black, 2-back red, 3-back blue, 4-back green) to avoid anchor effects in the ED procedure. To assess perceived task load, we will use the 6-item NASA Task Load Index (NASA-TLX),²⁶ 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. After each level, participants fill out the NASA-TLX on a tablet. Then, they complete the ED procedure on screen, where each possible pairing of the four n-back levels is presented in a randomized order. Participants are instructed to decide as realistically as possible, because one of their choices from the last iteration steps will be randomly chosen for one final run of n-back. This is only done to incentivise truthful behavior in the ED procedure, so the n-back data of this part will not be analyzed. The second session consists 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 follows the same structure of task and ED procedure, but participants can decide which strategy they want to reapply in the last block. Participants will receive 30€ in total or course credit for participation. Study data will be collected and managed using REDCap electronic data capture tools hosted at Technische Universität Dresden.^{27,28}

Sampling plan

A sample size analysis with *G×Power*,^{29,30} based on the results of the ANOVA of Westbrook et al.⁷ which showed an increase in reaction time with higher n-back levels, indicated that we should collect data from at least 53 participants, assuming $\eta^2 = 0.04$, $\alpha = .05$, and $\beta = .95$. The power analyses of all other hypotheses yielded smaller necessary sample sizes. To account for technical errors and exclusions of physiological data of the second lab session due to excessive noise, we aim to collect data of 60 to 70 participants.

Analysis plan

Data collection and analysis will not be performed blind to the conditions of the experiments. We aim to conduct all analysis as described in Westbrook et al.,⁷ but the level of detail was not always sufficient, so there might be deviations regarding data cleaning and degrees of freedom. The performance measure d' will be computed as the difference of the z -transformed hit rate and the z -transformed false alarm rate.³¹ Reaction time (RT) data will be trimmed by excluding all trials with responses faster than 100 ms, as the relevant cognitive processes cannot have been completed before.^{32,33} Aggregated RT values will be described using the median and the median of absolute deviation (*MAD*) as robust estimates of center and variability, respectively.³⁴ Error- and post-error trials will be excluded in repeated measures analyses of variance (rmANOVA) and controlled for in multi-level-model (MLM), because RT on the latter is longer due to more cautious behavior.^{35,36} To test our hypotheses, we will perform a series of rmANOVAs and an MLM

with orthogonal sum-to-zero contrasts in order to meaningfully interpret results.³⁷ Declining performance will be investigated by calculating an rmANOVA with three paired contrasts comparing d' between two levels of 2-, 3-, and 4-back at a time. Another rmANOVA with three paired contrasts will be computed to compare the mean RT between two levels of 2-, 3-, and 4-back at a time. To investigate changes in NASA-TLX ratings, six rmANOVAs will be computed, one for each NASA-TLX subscale, and each with six paired contrasts comparing the ratings between two levels of 1-, 2-, 3-, and 4-back at a time. For each ED round, SVs will be calculated by adding or subtracting 0.015625 from the last monetary value of the flexible level, depending on the participant's last choice. Then, these final monetary values will be divided by 2€, and the SV of each level per participant will be computed by averaging all final values of each level, regardless of whether it was fixed or flexible. An rmANOVA with six paired contrasts will be computed, comparing the SVs between two levels of 1-, 2-, 3-, and 4-back at a time. Tukey method will be used for the paired contrasts of each rmANOVA, including p -value adjustment.

To determine the influence of task performance in the association of SVs and n-back level, we will set up a MLM using the *lmerTest* package. We will apply restricted maximum likelihood (REML) to fit the model. First, we will calculate the intraclass correlation (ICC) on the basis of the null model. Second, we will estimate a random slopes model of SVs including n-back load level as level-1-predictor and, additionally, NFC as level-2-predictor. Within the model, we will control for d' , RT, correct, and post-correct trials.

$$SV \sim level * NFC + d' + RT + correct + postcorrect + (level|subject)$$

Level-1-predictors will be centered within cluster, whereas the level-2-predictor will be centered at the grand mean as recommended by Enders & Tofighi.³⁸ We will visually inspect the residuals of the final model. The approximately normal distribution indicates

no evidence to perform model criticism.

Third, we will perform a simple slopes analysis with n-back level as predictor and NFC as moderator. To evaluate the moderating effect, we will calculate the Johnson-Neyman interval.

To ensure the validity of the MLM, we will conduct a specification curve analysis,³⁹ which will include 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 MLM will be run with each of the 63 pipelines, which will also include our main pipeline (untransformed data, exclusion of RTs below 100 ms). The ratio of pipelines that lead to significant versus non-significant effects will provide an indication of how robust the effect actually is.

The association of ED and NFC will be examined with a regression using the *AUC* of each participant's SVs to predict their NFC score. A second regression will additionally include the mean of the NASA-TLX subscales' *AUCs* of each participant as a predictor. Since we do not have a fixed SV of 1 for 1-back, we cannot apply the “*AUC*” computation of Westbrook et al.,⁷ which was the mean of the *AUCs* of the SVs of each higher n-back level and 1-back, yielding values between 0 and 1. Consequently, we will choose a different way of quantifying the individual degree of ED. A classic *AUC* cannot differentiate between a subject who prefers 1-back and a subject who prefers 4-back if the magnitude of the ascent is the same, but it can reflect the overall willingness to exert effort. This is the opposite for the sum of the ascent between SVs. Therefore, we multiply both indicators, arriving at a value reflecting both degree and direction of preference, called *AxAUC*.

The results of each analysis will be assessed on the basis of both *p*-value and the Bayes factor *BF*10, calculated using the *BayesFactor* package.⁴⁰

243 **Data availability**

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

245 **Code availability**

246 The paradigm code, as well as the R Markdown file used to analyze the data and
247 write this document is available at our Github repository.

References

1.

Botvinick, M. M., Huffstetler, S. & McGuire, J. T. Effort discounting in human nucleus accumbens. *Cognitive, affective & behavioral neuroscience* **9**, 16–27 (2009).

2.

Kool, W. & Botvinick, M. Mental labour. *Nature Human Behaviour* **2**, 899–908 (2018).

3.

Mackworth, J. F. Paced memorizing in a continuous task. *Journal of Experimental Psychology* **58**, 206–211 (1959).

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).

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).

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).

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).

270 8.

271 Cacioppo, J. T. & Petty, R. E. The Need for Cognition. *Journal of Personality and*
272 *Social Psychology* **42**, 116–131 (1982).

273 9.

274 Wu, R., Ferguson, A. & Inzlicht, M. *Do humans prefer cognitive effort over doing*
275 *nothing?* <https://psyarxiv.com/d2gkf/> (2021) doi:10.31234/osf.io/d2gkf.

276 10.

277 Bertrams, A. & Dickhäuser, O. Passionate thinkers feel better. *Journal of Individual*
278 *Differences* **33**, 69–75 (2012).

279 11.

280 Nishiguchi, Y., Takano, K. & Tanno, Y. The Need for Cognition mediates and mod-
erates the association between depressive symptoms and impaired Effortful Control.
281 *Psychiatry Research* **241**, 8–13 (2016).

282 12.

283 Xu, P. & Cheng, J. Individual differences in social distancing and mask-wearing in the
pandemic of COVID-19: The role of need for cognition, self-control and risk attitude.
284 *Personality and Individual Differences* **175**, 110706 (2021).

285 13.

286 Kramer, A.-W., Van Duijvenvoorde, A. C. K., Krabbendam, L. & Huizenga, H. M.
Individual differences in adolescents' willingness to invest cognitive effort: Relation
to need for cognition, motivation and cognitive capacity. *Cognitive Development* **57**,
287 100978 (2021).

288 14.

Crawford, J. L., Eisenstein, S. A., Peelle, J. E. & Braver, T. S. Domain-general cognitive motivation: Evidence from economic decision-making. *Cognitive Research: Principles and Implications* **6**, 4 (2021).

15.

Culbreth, A., Westbrook, A. & Barch, D. Negative symptoms are associated with an increased subjective cost of cognitive effort. *Journal of Abnormal Psychology* **125**, 528–536 (2016).

16.

Westbrook, A., Lamichhane, B. & Braver, T. The subjective value of cognitive effort is encoded by a domain-general valuation network. *The Journal of Neuroscience* **39**, 3934–3947 (2019).

17.

Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* **51**, 195–203 (2019).

18.

R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2020).

19.

RStudio Team. RStudio: Integrated development for R. (2020).

20.

Singmann, H., Bolker, B., Westfall, J., Aust, F. & Ben-Shachar, M. S. *Afex: Analysis of factorial experiments*. (2021).

21.

310 Makowski, D., Ben-Shachar, M. S. & Lüdtke, D. bayestestR: Describing effects and
their uncertainty, existence and significance within the bayesian framework. *Journal*
311 *of Open Source Software* **4**, 1541 (2019).

312 22.

313 Greiner, B. Subject pool recruitment procedures: Organizing experiments with
314 ORSEE. *Journal of the Economic Science Association* **1**, 114–125 (2015).

315 23.

316 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*
317 *für Sozialpsychologie* **25**, (1994).

318 24.

319 Cacioppo, J. T., Petty, R. E. & Kao, C. F. The Efficient Assessment of Need for
320 Cognition. *Journal of Personality Assessment* **48**, 306–307 (1984).

321 25.

322 Fleischhauer, M. *et al.* Same or different? Clarifying the relationship of need for
cognition to personality and intelligence. *Personality & Social Psychology Bulletin*
323 **36**, 82–96 (2010).

324 26.

325 Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): Re-
326 sults of empirical and theoretical research. **52**, 139–183 (1988).

327 27.

328 Harris, P. A. *et al.* Research electronic data capture (REDCap)—A metadata-driven
methodology and workflow process for providing translational research informatics
329 support. *Journal of Biomedical Informatics* **42**, 377–381 (2009).

28.

Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics* **95**, 103208 (2019).

29.

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).

30.

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 Methods* **41**, 1149–1160 (2009).

31.

Macmillan, N. A. & Creelman, C. D. Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin* **107**, 401–413 (1990).

32.

Whelan, R. Effective Analysis of Reaction Time Data. *The Psychological Record* **58**, 475–482 (2008).

33.

Berger, A. & Kiefer, M. Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study. *Frontiers in Psychology* **12**, 2194 (2021).

34.

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 Psycholinguistics* **32**, 389–416 (2011).

351 35.

352 Dutilh, G. *et al.* Testing theories of post-error slowing. *Attention, Perception, &*
353 *Psychophysics* **74**, 454–465 (2012).

354 36.

355 Houtman, F., Castellar, E. N. & Notebaert, W. Orienting to errors with and without
356 immediate feedback. *Journal of Cognitive Psychology* **24**, 278–285 (2012).

357 37.

358 Singmann, H. & Kellen, D. An introduction to mixed models for experimen-
tal psychology. in *New methods in cognitive psychology* 4–31 (Routledge, 2019).
359 doi:10.4324/9780429318405-2.

360 38.

361 Enders, C. K. & Tofighi, D. Centering predictor variables in cross-sectional multilevel
362 models: A new look at an old issue. *Psychological Methods* **12**, 121–138 (2007).

363 39.

364 Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nature*
365 *Human Behaviour* **4**, 1208–1214 (2020).

366 40.

367 Morey, R. D. & Rouder, J. N. *BayesFactor: Computation of Bayes factors for common*
368 *designs*. (2021).

Acknowledgements

This research was partly funded by the German Research Foundation (DFG) as part of the Collaborative Research Center (CRC) 940. The funders have/had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. **OR** The authors received no specific funding for this work.

Author Contributions

JZ and CS conceptualized the study and its methodology, acquired funding, investigated, administered the project, and wrote the software. JZ and CK did the formal analysis, visualized the results, and prepared the original draft. All authors reviewed, edited, and approved the final version of the manuscript.

Competing Interests

The authors declare no competing interests.

Figures and figure Captions

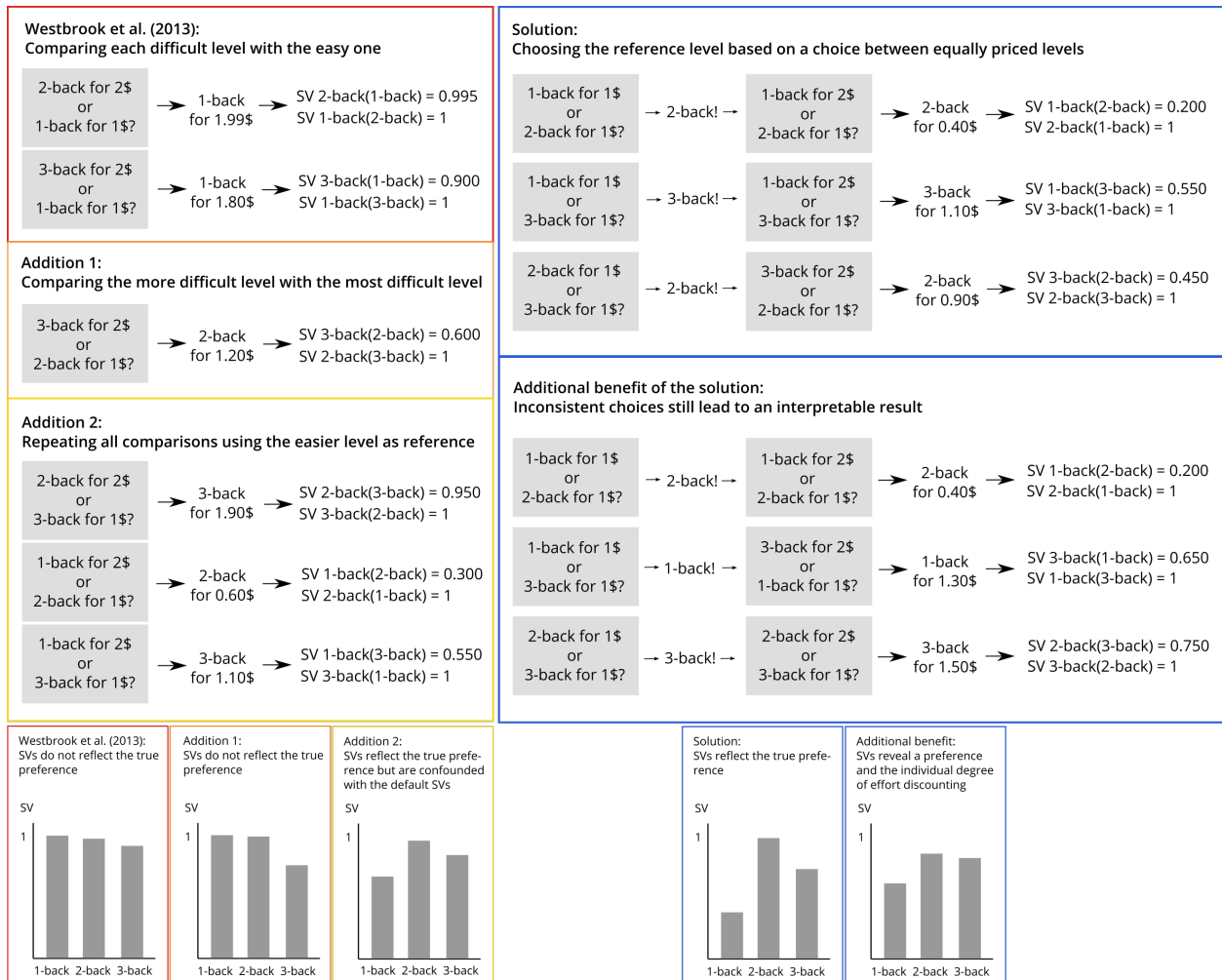


Figure 2. An example for subjective values for an n-back task with three levels, returned by different modifications of the COG-ED paradigm for a participant with the true preference 2-back > 3-back > 1-back.

382

Design Table

383

(Starts on next page)

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis Plan	Interpretation given to different outcomes
1. Do objective and subjective measures of performance reflect an increase in task load with increasing n-back level?	1a) The signal detection measure d' declines with increasing n-back level.	<p>F tests - ANOVA: Repeated measures, within factors</p> <p>Analysis: A priori: Compute required sample size</p> <p><u>Input:</u></p> <p>Effect size $f = 0.8685540$</p> <p>α err prob = 0.05</p> <p>Power ($1-\beta$ err prob) = 0.95</p> <p>Number of groups = 1</p> <p>Number of measurements = 4</p> <p>Corr among rep measures = 0.5</p> <p>Nonsphericity correction $\epsilon = 1$</p> <p><u>Output:</u></p> <p>Noncentrality parameter $\lambda = 30.1754420$</p> <p>Critical F = 3.4902948</p> <p>Numerator df = 3.0000000</p> <p>Denominator df = 12.0000000</p> <p>Total sample size = 5</p> <p>Actual power = 0.9824202</p>	<p>Repeated measures ANOVA with three linear contrasts, comparing the d' value of two n-back levels (2, 3, 4) at a time.</p> <p>The ANOVA is calculated using <code>aov_ez()</code> of the <code>afex</code>-package, estimated marginal means are calculated using <code>emmeans()</code> from the <code>emmeans</code>-package, and pairwise contrasts are calculated using <code>pairs()</code>.</p> <p>Bayes factors are computed for the ANOVA and each contrast using the <code>BayesFactor</code>-package.</p>	<p>ANOVA yields $p < .05$ is interpreted as d' changing significantly with n-back levels. Values of d' are interpreted as equal between n-back levels if $p > .05$.</p> <p>Each contrast yielding $p < .05$ is interpreted as d' being different between those levels, magnitude and direction are inferred from the respective estimate. Values of d' are interpreted as equal between n-back levels if $p > .05$.</p> <p>The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.</p>
	1b) Reaction time increases with increasing n-back level.	<p>F tests - ANOVA: Repeated measures, within factors</p> <p>Analysis: A priori: Compute required sample size</p> <p><u>Input:</u></p> <p>Effect size $f = 0.2041241$</p> <p>α err prob = 0.05</p> <p>Power ($1-\beta$ err prob) = 0.95</p> <p>Number of groups = 1</p>	<p>Repeated measures ANOVA with three linear contrasts, comparing the median reaction time of two n-back levels (2, 3, 4) at a time.</p> <p>The ANOVA is calculated using <code>aov_ez()</code> of the <code>afex</code>-package, estimated marginal means are</p>	<p>ANOVA yields $p < .05$ is interpreted as the median reaction time changing significantly with n-back levels. Median reaction times are interpreted as equal between n-back levels if $p > .05$.</p>

		<p>Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction $\epsilon = 1$ <u>Output:</u> Noncentrality parameter $\lambda = 17.6666588$ Critical F = 2.6625685 Numerator df = 3.0000000 Denominator df = 156 Total sample size = 53 Actual power = 0.9506921</p>	<p>calculated using emmeans() from the emmeans-package, and pairwise contrasts are calculated using pairs().</p> <p>Bayes factors are computed for the ANOVA and each contrast using the BayesFactor-package.</p>	<p>Each contrast yielding $p < .05$ is interpreted as the median reaction time being different between those levels, magnitude and direction are inferred from the respective estimate. Median reaction times are interpreted as equal between n-back levels if $p > .05$.</p> <p>The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.</p>
	1c) Ratings on all NTLX subscales increase with increasing n-back level.	<p>From Kramer et al.: F tests - ANOVA: Repeated measures, within factors Analysis: A priori: Compute required sample size <u>Input:</u> Effect size $f = 0.7071068$ α err prob = 0.05 Power ($1 - \beta$ err prob) = 0.95 Number of groups = 1 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction $\epsilon = 1$ <u>Output:</u> Noncentrality parameter $\lambda = 24.0000013$ Critical F = 3.2873821</p>	<p>A repeated measures ANOVA for each NASA-TLX subscale, with six linear contrasts comparing the subscale score of two n-back levels (1, 2, 3, 4) at a time.</p> <p>The ANOVA is calculated using aov_ez() of the afex-package, estimated marginal means are calculated using emmeans() from the emmeans-package, and pairwise contrasts are calculated using pairs().</p>	<p>ANOVA yields $p < .05$ is interpreted as the subscale score changing significantly with n-back levels. The subscale scores are interpreted as equal between n-back levels if $p > .05$.</p> <p>Each contrast yielding $p < .05$ is interpreted as the subscale score being different between those levels, magnitude and direction are inferred from the respective estimate. The subscale scores are interpreted</p>

		<p>Numerator df = 3.0000000 Denominator df = 15.0000000 Total sample size = 6 Actual power = 0.9620526</p>	<p>Bayes factors are computed for the ANOVA and each contrast using the BayesFactor-package.</p>	<p>as equal between n-back levels if $p > .05$.</p> <p>The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.</p>
<p>2. Is the effort required for higher n-back levels less attractive, regardless of how well a person performs?</p>	<p>2a) Subjective values decline with increasing n-back level.</p>	<p>F tests - ANOVA: Repeated measures, within factors Analysis: A priori: Compute required sample size <u>Input:</u> Effect size $f = 0.9229582$ α err prob = 0.05 Power ($1 - \beta$ err prob) = 0.95 Number of groups = 1 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction $\epsilon = 1$ <u>Output:</u> Noncentrality parameter $\lambda = 27.2592588$ Critical F = 3.8625484 Numerator df = 3.0000000 Denominator df = 9.0000000 Total sample size = 4 Actual power = 0.9506771</p>	<p>Repeated measures ANOVA with six linear contrasts, comparing the subjective values of two n-back levels (1, 2, 3, 4) at a time.</p> <p>The ANOVA is calculated using <code>aov_ez()</code> of the afex-package, estimated marginal means are calculated using <code>emmeans()</code> from the emmeans-package, and pairwise contrasts are calculated using <code>pairs()</code>.</p> <p>Bayes factors are computed for the ANOVA and each contrast using the BayesFactor-package.</p>	<p>ANOVA yields $p < .05$ is interpreted as subjective values changing significantly with n-back levels. Subjective values are interpreted as equal between n-back levels if $p > .05$.</p> <p>Each contrast yielding $p < .05$ is interpreted as subjective values being different between those levels, magnitude and direction are inferred from the respective estimate. Subjective values are interpreted as equal between n-back levels if $p > .05$.</p> <p>The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.</p>

	2b) Subjective values decline with increasing n-back level, even after controlling for declining task performance measured by signal detection d' and reaction time.	<p>t tests - Linear multiple regression: Fixed model, single regression coefficient Analysis: A priori: Compute required sample size <u>Input:</u> Tail(s) = One Effect size $f^2 = 0.34$ α err prob = 0.05 Power ($1-\beta$ err prob) = 0.95 Number of predictors = 3 <u>Output:</u> Noncentrality parameter $\delta = 3.4000000$ Critical t = 1.6955188 Df = 31 Total sample size = 34 Actual power = 0.9534767</p>	<p>[Cursive refers to 2c] Multilevel model of SVs with n-back load level as level-1-predictor <i>and NFC as level-2-predictor</i> controlling for d', reaction time, correct and post-correct trials using subject-specific intercepts and allowing random slopes for n-back level.</p> <p>The null model and the random slopes model are calculated using lmer() of the lmerTest-package. <i>Simple slopes analysis and Johnson-Neyman intervals are performed using the functions sim_slopes() and johnson_neyman() of the interactions-package.</i></p> <p>Bayes factors are computed for the MLM using the BayesFactor-package.</p>	<p>[Cursive refers to 2c] Fixed effects yield $p < .05$ are interpreted as subjective values changing significantly with n-back levels <i>and NFC-score, respectively</i>. Subjective values are interpreted as equal between n-back levels if $p > .05$.</p> <p><i>Simple slopes of level for values of NFC yield $p < .05$ are interpreted as subjective values changing significantly with n-back levels for the specific value of NFC. Subjective values are interpreted as equal between n-back levels for specific values of NFC if $p > .05$.</i></p> <p>The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.</p>
	2c) SVs decline stronger with increasing task load for individuals with low compared to high NFC scores.			
3. Is there a discrepancy between perceived task load and subjective value of effort	3a) Subjective values positively predict individual NFC scores.	<p>t tests - Linear multiple regression: Fixed model, single regression coefficient Analysis: A priori: Compute required sample size <u>Input:</u> Tail(s) = One Effect size $f^2 = 0.33$</p>	Subjective values are regressed on NFC scores using the lm() function from the stats-package.	Subjective values are interpreted as predicting NFC scores if the slope yields $p < .05$. Direction and magnitude are inferred from the slope estimate.

depending on a person's Need for Cognition?		α err prob = 0.05 Power ($1-\beta$ err prob) = 0.95 Number of predictors = 1 <u>Output:</u> Noncentrality parameter $\delta = 3.3985291$ Critical t = 1.6923603 Df = 33 Total sample size = 35 Actual power = 0.9537894	Bayes factors are computed for the regression using the BayesFactor-package.	The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.
	3b) NASA-TLX scores negatively predict individual NFC scores.	Westbrook et al. have only reported the p-value here, so we used the regression results of our pilot study, which included NASA-TLX scores and subjective values as predictors of NFC scores. t tests - Linear multiple regression: Fixed model, single regression coefficient Analysis: A priori: Compute required sample size <u>Input:</u> Tail(s) = One Effect size $f^2 = 1.10$ α err prob = 0.05 Power ($1-\beta$ err prob) = 0.95 Number of predictors = 2 <u>Output:</u> Noncentrality parameter $\delta = 3.6331804$ Critical t = 1.8331129 Df = 9 Total sample size = 12 Actual power = 0.9552071	Subjective values and the area under the curve of each subject's NASA-TLX scores are regressed on NFC scores using the lm() function from the stats-package. Bayes factors are computed for each predictor using the BayesFactor-package.	Subjective values and NASA-TLX scores are interpreted as predicting NFC scores if their slope yields $p < .05$. Direction and magnitude are inferred from the slope estimate. The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.

Supplement

Results of the pilot study

Hypothesis 1a: The signal detection measure d' declines with increasing n-back level.

ANOVA:

$$F(1.86, 26.06) = 0.00, MSE = 1.67, p > .999, \eta_p^2 = 1.43\text{e-}32, 95\% \text{ CI } [[0.00, 1.00]]$$

Paired contrasts:

Table 1

Paired contrasts for the rmANOVA comparing d' between n-back levels

Contrast	Estimate	SE	df	t	p	BF_{10}
2 - 3	0.00	0.46	28.00	0.00	1.00	0.26
2 - 4	0.00	0.46	28.00	0.00	1.00	0.26
3 - 4	0.00	0.46	28.00	0.00	1.00	0.26

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.

Hypothesis 1b: Reaction time increases with increasing n-back level.

ANOVA:

$$F(1.76, 24.71) = 5.59, MSE = 0.01, p = .012, \hat{\eta}_G^2 = .077$$

Paired contrasts:

Table 2

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

Contrast	Estimate	SE	df	t	p	BF10
2 - 3	-0.10	0.03	28.00	-3.24	0.01	8.45
2 - 4	-0.03	0.03	28.00	-0.89	0.65	0.34
3 - 4	0.08	0.03	28.00	2.35	0.07	4.49

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.

Hypothesis 1c: Ratings on all NASA-TLX dimensions increase with increasing n-back level.

Mental subscale ANOVA:

$F(2.08, 27.03) = 69.96$, $MSE = 6.47$, $p < .001$, $\hat{\eta}_G^2 = .628$, $BF10 = 240,305,851.21$

Mental subscale paired contrasts:

Table 3

Paired contrasts for the rmANOVA comparing ratings on the NASA-TLX Mental subscale between n-back levels

Contrast	Estimate	SE	df	t	p	BF10
1 - 2	-4.43	0.80	39.00	-5.53	0.00	1,400.60
1 - 3	-8.43	0.80	39.00	-10.53	0.00	35,718.31
1 - 4	-10.79	0.80	39.00	-13.47	0.00	189,999.47
2 - 3	-4.00	0.80	39.00	-5.00	0.00	372.90
2 - 4	-6.36	0.80	39.00	-7.94	0.00	3,326.17
3 - 4	-2.36	0.80	39.00	-2.94	0.03	38.13

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.

Physical subscale ANOVA:

$$F(1.61, 20.96) = 7.86, MSE = 8.31, p = .005, \hat{\eta}_G^2 = .071, BF10 = 0.34$$

Physical subscale paired contrasts:

Table 4

Paired contrasts for the rmANOVA comparing ratings on the NASA-TLX Physical subscale between n-back levels

Contrast	Estimate	SE	df	t	p	BF10
1 - 2	-1.64	0.80	39.00	-2.06	0.19	3.51
1 - 3	-3.07	0.80	39.00	-3.85	0.00	6.50
1 - 4	-3.50	0.80	39.00	-4.38	0.00	7.66
2 - 3	-1.43	0.80	39.00	-1.79	0.29	1.79
2 - 4	-1.86	0.80	39.00	-2.33	0.11	2.00
3 - 4	-0.43	0.80	39.00	-0.54	0.95	0.38

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.

Time subscale ANOVA:

$$F(2.14, 27.86) = 31.25, MSE = 6.62, p < .001, \hat{\eta}_G^2 = .254, BF10 = 24.80$$

Time subscale paired contrasts:

Table 5

Paired contrasts for the rmANOVA comparing ratings on the NASA-TLX Time subscale between n-back levels

Contrast	Estimate	SE	df	t	p	BF10
1 - 2	-1.64	0.82	39.00	-2.00	0.21	11.44
1 - 3	-5.14	0.82	39.00	-6.26	0.00	278.18
1 - 4	-7.14	0.82	39.00	-8.69	0.00	3,713.67
2 - 3	-3.50	0.82	39.00	-4.26	0.00	38.79
2 - 4	-5.50	0.82	39.00	-6.69	0.00	1,064.28
3 - 4	-2.00	0.82	39.00	-2.43	0.09	3.09

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.

Performance subscale ANOVA:

$$F(2.12, 27.59) = 6.78, MSE = 11.87, p = .004, \hat{\eta}_G^2 = .151, BF10 = 1.82$$

Performance subscale paired contrasts:

Table 6

Paired contrasts for the rmANOVA comparing ratings on the NASA-TLX Performance subscale between n-back levels

Contrast	Estimate	SE	df	t	p	BF10
1 - 2	1.50	1.10	39.00	1.37	0.53	1.00
1 - 3	3.93	1.10	39.00	3.59	0.00	33.72
1 - 4	4.21	1.10	39.00	3.85	0.00	5.32
2 - 3	2.43	1.10	39.00	2.22	0.14	10.97
2 - 4	2.71	1.10	39.00	2.48	0.08	1.83
3 - 4	0.29	1.10	39.00	0.26	0.99	0.28

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.

Effort subscale ANOVA:

$$F(1.57, 20.43) = 28.65, MSE = 12.23, p < .001, \hat{\eta}_G^2 = .433, BF10 = 10,733.57$$

Effort subscale paired contrasts:

Table 7

Paired contrasts for the rmANOVA comparing ratings on the NASA-TLX Effort subscale between n-back levels

Contrast	Estimate	SE	df	t	p	BF10
1 - 2	-2.71	0.96	39.00	-2.84	0.03	1,015.57
1 - 3	-6.79	0.96	39.00	-7.09	0.00	774.36
1 - 4	-7.79	0.96	39.00	-8.14	0.00	1,383.62
2 - 3	-4.07	0.96	39.00	-4.26	0.00	55.57
2 - 4	-5.07	0.96	39.00	-5.30	0.00	44.55
3 - 4	-1.00	0.96	39.00	-1.05	0.72	0.62

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.

Frustration subscale ANOVA:

$F(2.53, 32.94) = 35.31$, $MSE = 6.85$, $p < .001$, $\hat{\eta}_G^2 = .445$, $BF10 = 17,679.16$

Frustration subscale paired contrasts:

Table 8

Paired contrasts for the rmANOVA comparing ratings on the NASA-TLX Frustration subscale between n-back levels

Contrast	Estimate	SE	df	t	p	BF10
1 - 2	-1.57	0.91	39.00	-1.73	0.32	3.52
1 - 3	-5.71	0.91	39.00	-6.28	0.00	589.81
1 - 4	-8.36	0.91	39.00	-9.19	0.00	27,016.64
2 - 3	-4.14	0.91	39.00	-4.56	0.00	71.13
2 - 4	-6.79	0.91	39.00	-7.46	0.00	2,658.32
3 - 4	-2.64	0.91	39.00	-2.91	0.03	2.54

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.

Hypothesis 2a: Subjective values decline with increasing n-back level.

ANOVA:

$F(1.80, 25.26) = 7.80$, $MSE = 0.06$, $p = .003$, $\hat{\eta}_G^2 = .269$, $BF10 = 62.57$

Paired contrasts:

Table 9

Paired contrasts for the rmANOVA comparing subjective values between n-back levels

Contrast	Estimate	SE	df	t	p	BF10
1 - 2	0.08	0.07	42.00	1.12	0.68	0.65
1 - 3	0.17	0.07	42.00	2.46	0.08	4.65
1 - 4	0.32	0.07	42.00	4.59	0.00	7.97
2 - 3	0.09	0.07	42.00	1.34	0.54	1.18
2 - 4	0.24	0.07	42.00	3.48	0.01	17.86
3 - 4	0.15	0.07	42.00	2.13	0.16	1.08

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.

Hypothesis 2b: Subjective values decline with increasing n-back level, even after controlling for declining task performance measured by signal detection d' and reaction time.

Multi level model:

Table 10

Effects of n-back load level on subjective value controlled for task performance (d' and reaction time), correct and postcorrect trials.

Parameter	Beta	SE	p-value	Random Effects (SD)
Intercept	0.75	0.05	<.001***	0.18
n-back level	-0.12	0.04	0.005**	0.14
NFC	0.00	0.01	0.906	
d'	0.04	0.00	<.001***	
RT	0.04	0.01	<.001***	
level x NFC	0.00	0.00	0.38	

Note: NFC = Need for Cognition, SE = standard error.

*** $p < .001$, ** $p < .01$, * $p < 0.5$.

Hypothesis 2c: Subjective values decline stronger with increasing task load for individuals with low compared to high NFC scores.

Simple slopes analysis:

Table 11

Interaction between NFC and n-back load level.

Value of NFC	Slopes of NFC				Conditional Intercept	
	Beta	SE	95% CI	p-value	Beta	SE
- 1 SD	-0.09	0.05	[-0.19,0.01]	.098	0.76	0.07
Mean	-0.12	0.04	[-0.19,-0.05]	.005**	0.75	0.05
+ 1 SD	-0.16	0.05	[-0.26,-0.06]	.009**	0.75	0.07

Note: NFC = Need for Cognition, SE = standard error. *** $p < .001$, ** $p < .01$, * $p < 0.5$.

Johnson-Neyman intervals:

-6.97 and 21.76

Specification curve analysis:

Hypothesis 3a: Subjective values positively predict individual NCS scores.

Intercept: $b = 20.65$, 95% CI [13.19, 28.11]

Predictor $AxAUC$: $b = -1.41$, 95% CI [-8.20, 5.37]

Fit: $R^2 = .02$, 90% CI [0.00, 0.27]

$BF_{10} = 0.51$

Hypothesis 3b: NASA-TLX scores negatively predict individual NFC scores.

Intercept: $b = 39.56$, 95% CI [26.20, 52.92]

Predictor $AxAUC$: $b = -4.04$, 95% CI [-9.31, 1.22]

Predictor AUC NASA-TLX: $b = -0.71$, 95% CI [-1.16, -0.25]

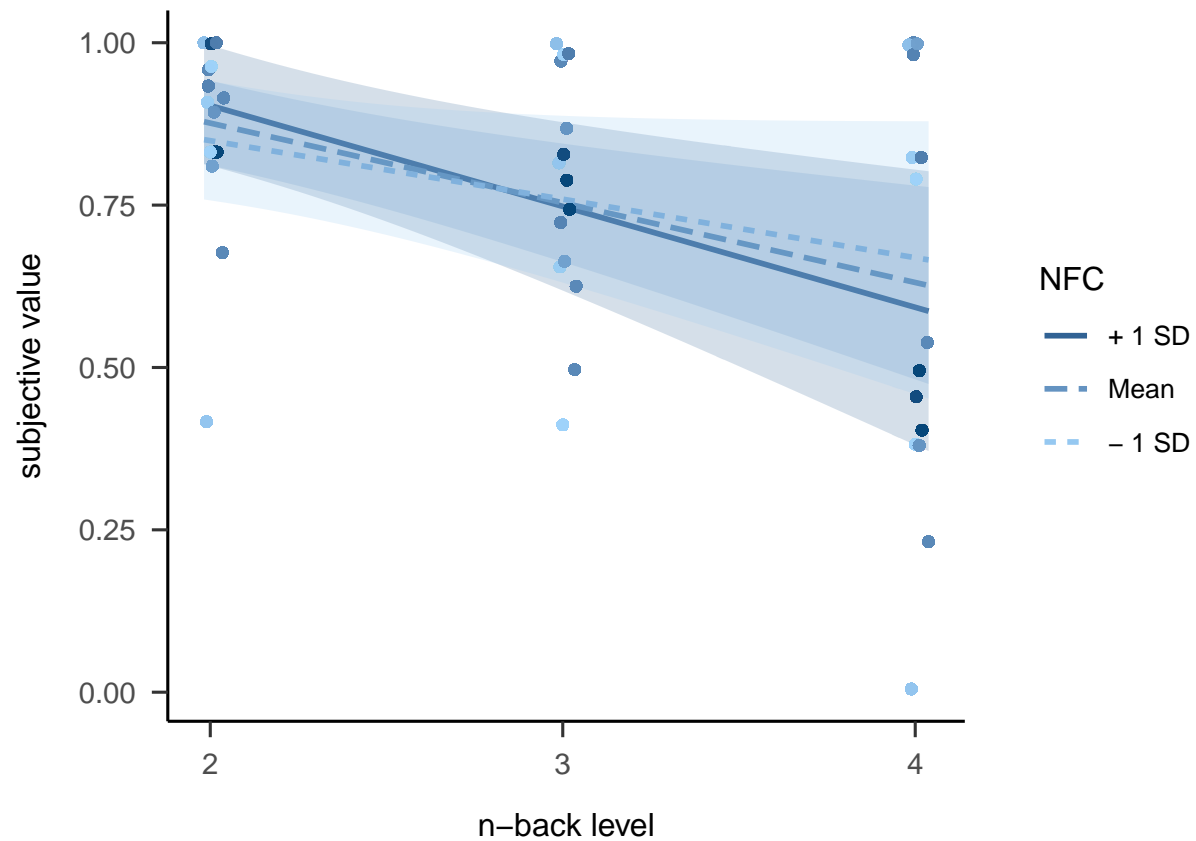


Figure 3. Simple slopes analysis for how n-back level predicts the subjective value, depending on the participant's NFC. NFC = Need for Cognition, SD = standard deviation.

Fit: $R^2 = .52$, 90% CI [0.08, 0.75]

$BF_{10} = 0.48$, 3.88, and 4.17

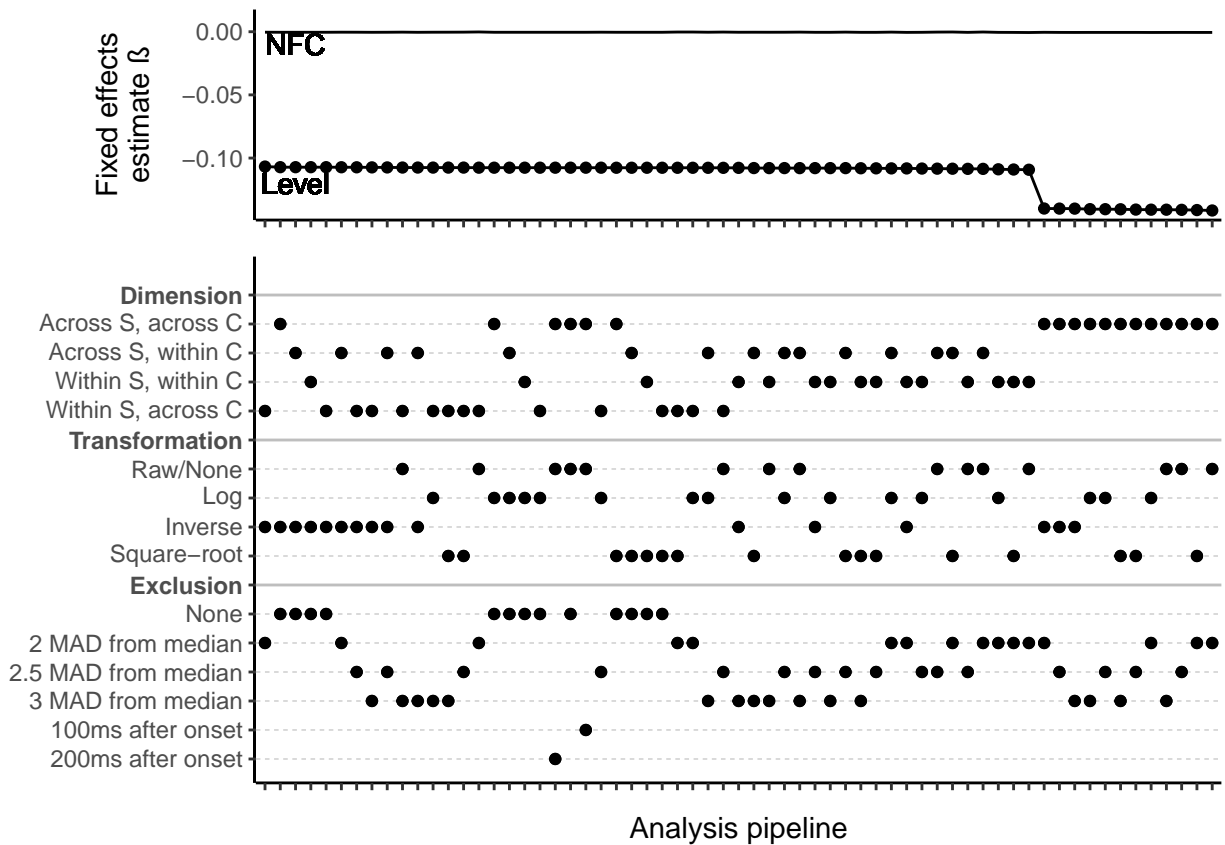


Figure 4. Results of the specification curve analysis for the multi level model. The upper panel shows the fixed effect estimates for Need for Cognition and n-back level as predictors of subjective values. Estimates with $p < .01$ are indicated by a dot on the line. The lower panel shows the preprocessing steps of each corresponding pipeline. The BF_{10} of each pipeline's multi level model approached infinity.