The effect sizes for each hypothesis were taken from the corresponding analysis in Westbrook et al. (2013). There are two exceptions due to the fact that the information in Westbrook et al. (2013) was insufficient in that case: Hypothesis 1c was based on Kramer et al. (2021), and hypothesis 3b was based on our pilot data.

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.	F tests - ANOVA: Repeated measures, within factors Analysis: A priori: Compute required sample size <u>Input</u>: Effect size f = 0.8685540 α err prob = 0.05 Power (1-β err prob) = 0.95 Number of groups = 1 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction ε = 1 <u>Output</u>: Noncentrality parameter λ = 30.1754420 Critical F = 3.4902948 Numerator df = 3.0000000 Denominator df = 12.0000000 Total sample size = 5 Actual power = 0.9824202	Repeated measures ANOVA with six linear contrasts, comparing the d' value of two n-back levels (1, 2, 3, 4) at a time. 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(). Bayes factors are computed for the ANOVA and each contrast using the BayesFactor-package.	ANOVA yields $p < .05$ is interpreted as d' changing significantly with n-back levels. Each contrast yielding $p < .05$ is interpreted as d' being different between those levels, magnitude and direction are inferred from the respective estimate. The Bayes factor *BF10* is reported alongside every *p*-value to assess the strength of evidence.
	1b) Reaction time increases with increasing n-back level.	F tests - ANOVA: Repeated measures, within factors Analysis: A priori: Compute required sample size <u>Input</u>: Effect size f = 0.2041241 α err prob = 0.05 Power (1-β err prob) = 0.95 Number of groups = 1 Number of measurements = 4	Repeated measures ANOVA with six linear contrasts, comparing the median reaction time of two n-back levels (1, 2, 3, 4) at a time. The ANOVA is calculated using aov_ez() of the afex-package, estimated marginal means are calculated using emmeans()	ANOVA yields $p < .05$ is interpreted as the median reaction time changing significantly with n-back levels. Each contrast yielding $p < .05$ is interpreted as the median reaction time being different between those levels, magnitude

		Corr among rep measures = 0.5 Nonsphericity correction ε = 1 <u>Output</u>: Noncentrality parameter λ = 17.6666588 Critical F = 2.6625685 Numerator df = 3.0000000 Denominator df = 156 Total sample size = 53 Actual power = 0.9506921	from the emmeans-package, and pairwise contrasts are calculated using pairs(). Bayes factors are computed for the ANOVA and each contrast using the BayesFactor-package.	and direction are inferred from the respective estimate. The Bayes factor *BF10* is reported alongside every *p*-value to assess the strength of evidence.
	1c) Ratings on all NTLX subscales increase with increasing n-back level.	From Kramer et al. (2021): 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-β err prob) = 0.95 Number of groups = 1 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction ε = 1 <u>Output</u>: Noncentrality parameter λ = 24.0000013 Critical F = 3.2873821 Numerator df = 3.0000000 Denominator df = 15.0000000 Total sample size = 6 Actual power = 0.9620526	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. 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(). Bayes factors are computed for the ANOVA and each contrast using the BayesFactor-package.	ANOVA yields $p < .05$ is interpreted as the subscale score changing significantly with n-back levels. 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 Bayes factor *BF10* is reported alongside every *p*-value to assess the strength of evidence.

2. Is the effort required for higher n-back levels less attractive, regardless of how well a person performs?	2a) Subjective values decline with increasing n-back level.	F tests - ANOVA: Repeated measures, within factors Analysis: A priori: Compute required sample size Input: Effect size f = 0.9229582 α err prob = 0.05 Power (1-β err prob) = 0.95 Number of groups = 1 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction ε = 1 Output: Noncentrality parameter λ = 27.2592588 Critical F = 3.8625484 Numerator df = 3.0000000 Denominator df = 9.0000000 Total sample size = 4 Actual power = 0.9506771	Repeated measures ANOVA with four contrasts (linear (3,1,-1,-3), quadratic (-1,1,1,-1), logistic (3,2,-2,-3), and skewed normal (1,2,-1,-2)), comparing the subjective values of all n-back levels. The ANOVA is calculated using aov_ez() of the afex-package, estimated marginal means are calculated using emmeans() from the emmeans-package. Bayes factors are computed for the ANOVA and each contrast using the BayesFactor-package.	ANOVA yields $p < .05$ is interpreted as subjective values changing significantly with n-back levels. Each contrast yielding $p < .05$ is interpreted as subjective values being different between levels, magnitude and direction are inferred from the respective estimate. The Bayes factor *BF10* is reported alongside every *p*-value to assess the strength of evidence.
	2b) Subjective values decline with increasing n-back level, even after controlling for declining task performance measured by signal detection d' and reaction time.	F tests - ANOVA: Repeated measures, within factors Analysis: A priori: Compute required sample size Input: Effect size f = 0.9229582 α err prob = 0.05 Power (1-β err prob) = 0.95 Number of groups = 1 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction ε = 1 Output: Noncentrality parameter λ = 27.2592588 Critical F = 3.8625484 Numerator df = 3.0000000	Multilevel model of SVs with n-back load level as level-1-predictor controlling for d' and reaction time subject-specific intercepts and allowing random slopes for n-back level. The null model and the random slopes model are calculated using lmer() of the lmerTest-package. Bayes factors are computed for the MLM using the BayesFactor-package.	Fixed effects yielding $p < .05$ are interpreted as subjective values changing significantly with n-back levels. The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.

		Denominator df = 9.0000000 Total sample size = 4 Actual power = 0.9506771		
3. Is there a discrepancy between perceived task load and subjective value of effort depending on a person's Need for Cognition?	3a) Participants with higher NFC scores have higher subjective values for 2- and 3-back but lower subjective values for 1-back than participants with lower NFC scores.	F tests - ANOVA: Repeated measures, within-between interaction: A priori: Compute required sample size <u>Input:</u> Effect size f = 0.57 α err prob = 0.05 Power (1-β err prob) = 0.95 Number of groups = 2 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction ε = 1 <u>Output:</u> Noncentrality parameter λ = 25.99 Critical F = 23.01 Numerator df = 3 Denominator df = 24 Total sample size = 10	Difference scores of subjective values are computed between consecutive n-back levels, and the sample is divided by their NFC median, so an rmANOVA with the within-factor n-back level and the between-factor NFC group can be computed. The ANOVA is calculated using aov_ez() of the afex-package, estimated marginal means are calculated using emmeans() from the emmeans-package. Bayes factors are computed for the ANOVA using the BayesFactor-package.	Subjective values are interpreted as being lower for 1-back and higher for 2- and 3-back in participants with higher NFC if there is a main effect of the NFC group ($p < .05$) and if the contrasts reveal that pattern at $p < .05$. The Bayes factor BF10 is reported alongside every *p*-value to assess the strength of evidence.
	3b) Participants with higher NFC scores have lower NASA-TLX scores in every n-back level than participants with lower NFC scores.	Westbrook et al. have only reported the p-value here, so we used the ANOVA results of our pilot study, which included NASA-TLX scores (per level and subject) and NFC scores. The F statistic was $F(1,12) = 7.57$, which is an effect size of $f = 0.7355$. F tests - ANOVA: Repeated measures, within-between interaction: A priori: Compute required sample size <u>Input:</u> Effect size f = 0.7355	NASA-TLX sum scores are computed per level and subject, and the sample is divided by their NFC median, so an rmANOVA with the within-factor n-back level and the between-factor NFC group can be computed. The ANOVA is calculated using aov_ez() of the afex-package, estimated marginal means are	NASA-TLX scores are interpreted as being lower for participants with higher NFC if there is a main effect of the NFC group ($p < .05$) and if the contrasts reveal that pattern at $p < .05$. The Bayes factor BF10 is reported alongside every p-value to assess the strength of evidence.

		α err prob = 0.05 Power (1-β err prob) = 0.95 Number of groups = 2 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction ε = 1 <u>Output:</u> Noncentrality parameter λ = 25.97 Critical F = 3.49 Numerator df = 3 Denominator df = 12 Total sample size = 6	calculated using emmeans() from the emmeans-package. Bayes factors are computed for each predictor using the BayesFactor-package.	
	3c) Participants with higher NFC scores have lower aversiveness ratings for 2- and 3-back but higher higher aversiveness ratings for 1-back than participants with lower NFC scores.	As we could not find any study reporting an association of NFC and aversiveness ratings, we assumed a medium to large association (r = 0.25, according to Gignac & Szodorai (2016), doi: 10.1016/j.paid.2016.06.069). We assume this, because NFC is a trait defined as a preference for effortful cognitive activities, thereby it should be negatively associated with aversion to a cognitively effortful task. F tests - ANOVA: Repeated measures, within-between interaction: A priori: Compute required sample size <u>Input:</u> Effect size f = 0.2582 α err prob = 0.05 Power (1-β err prob) = 0.95 Number of groups = 2 Number of measurements = 4 Corr among rep measures = 0.5 Nonsphericity correction ε = 1	Difference scores of aversiveness ratings are computed between consecutive n-back levels, and the sample is divided by their NFC median, so an rmANOVA with the within-factor n-back level and the between-factor NFC group can be computed. The ANOVA is calculated using aov_ez() of the afex-package, estimated marginal means are calculated using emmeans() from the emmeans-package. Bayes factors are computed for the ANOVA using the BayesFactor-package.	Aversiveness ratings are interpreted as being higher for 1-back and lower for 2- and 3-back in participants with higher NFC if there is a main effect of the NFC group ($p < .05$) and if the contrasts reveal that pattern at $p < .05$. The Bayes factor BF10 is reported alongside every *p*-value to assess the strength of evidence.

		Output: Noncentrality parameter $\lambda = 18.13$ Critical F = 2.70 Numerator df = 3 Denominator df = 96 Total sample size = 34		