UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Department of Psychology – Faculty of Health Sciences

Rationality: Can it be predicted by cognitive effort, ability and thinking disposition?

On the role of willingness to exert cognitive effort, thinking disposition and executive function on deliberate reasoning tasks both with and without a heuristic response.

Kristoffer Klevjer

Master's thesis in psychology – May, 2019

Sammendrag

Å bruke intuisjon og magefølelsen, når det som egentlig kreves er nøye resonnering kan føre til feilaktige vurdering og beslutninger, ikke bare på quiz, men også ha store konsekvenser i hverdagen. Så hvorfor gjør vi det? En vanlig forklaring er at vi gjøre det fordi det er strevsomt og krever innsats å bedrive nøye resonnering, innsats vi til vanlig ikke liker. In denne studien kastet vi lys over dette forholdet mellom (vellykket) resonnering og vilje til å yte kognitiv innsats.

Vi måte vilje til å yte kognitiv innsats ved å bruke to forskjellige eksperimentelle tilnærminger, samt en selvrapporteringsmåling. Og vi måte kritisk tenkning i et oppgavesett som både inneholdt spørsmål med sterke intuitive svar, og spørsmål uten intuitive svar. Alle oppgavene krevede nøye resonnering for å komme frem til det korrekte svaret, men oppgavene med intuitive svar krevde i tillegg at man ble oppmerksom på disse og unnlot å svare i henhold til de.

Våre eksperimentelle tilnærminger for å måle viljen til å yte kognitiv innsats viste seg mindre pålitelige, spesielt til bruk på individuelt nivå. Derimot så fant vi at vilje til å yte kognitiv innsats, målt gjennom selvrapporteringsskjema, og høyere kognitiv evne, målt via en arbeidshukommelsestest, førte til bedre skårer på kritisk tenkning.

Videre analyser indikerte derimot at dette hovedsakelig gjalt på oppgavene uten et sterkt intuitivt svar. Mens de aller fleste vil klare å utføre resonneringen som krevdes i oppgavene, synes den kritiske faktoren å være om man oppdager at resonnering kreves eller ikke, og dette ble ikke predikert ut i fra hverken vilje til å yte kognitiv innsats, eller kognitiv evne.

Nøkkelord: Rasjonalitet, kritisk tenkning, kognitiv innsats, heuristisk respons, resonnering

Acknowledgement

First and foremost I am profoundly grateful to my supervisor, Dr. Gerit Pfuhl at the Department of Psychology, UiT The Arctic University of Norway, for giving me the opportunity to do research and write my thesis within my field of interest, and providing continued support and encouragement throughout my time here. Your knowledge and expertise both within the field of research and research as a discipline never ceases to impress!

I would like to thank all my participants, providing me their valuable time and effort during long testing-sessions.

I would also like to thank the Department of Psychology, UiT, for providing such an open and flexible master's program, allowing me to choose which door to knock at when deciding upon a field and supervisor. Special thanks to the master committee, Tove, Frank, Sarah and Jon-André! With the latter always ready to help with any strange request or challenge I brought him.

I feel honored having been a part of this department, and hope to get the opportunity to work with you all again someday.

Lastly, I would like to express my deepest gratitude to my family, friends and close ones for their continued support (and understanding of my absence).

Thank you all! – Kristoffer, 28.04.2019

Kristoffer Klevjer

Kristoffer Klevjer

Gerit Pfuhl

Rationality: Can it be predicted by cognitive effort, ability and thinking disposition?

Kristoffer Klevjer

PSY-3900

Master's thesis in psychology

May, 2019

UiT The Arctic University of Norway

Abstract

Background

The use of intuitive responses when deliberate reasoning is needed leads to incorrect judgements and decisions, not only in quizzes, but might also impose large consequences in everyday life. A common explanation for this use of intuitive answers is due to the effort demands associated with deliberate reasoning. In this study we aimed to shed more light on the relationship between (successful) reasoning and willingness to exert cognitive effort.

Methods

We measured willingness to exert cognitive effort using two different experimental paradigms, as well as one self-report measurement. And we measured critical thinking in a task with items both with and without a prevalent intuitive answer. All of which required successful deliberate reasoning in order to reach the correct answer, however the intuitive items required a detection and suppression of the intuitive response as well.

Results

Our measures of willingness to exert cognitive effort proved less reliable, however critical thinking was increased with higher cognitive ability, as measured in an executive function measurement, and with a higher self-reported disposition towards complex thinking.

Conclusion

While critical thinking was modestly predicted by cognitive ability and disposition towards complex thinking, exploratory analyses indicated that this was less so in tasks with a strong intuitive response. While most individuals might be able to carry out the deliberate reasoning in these tasks, the critical factor seems to be whether or not they detect the need for performing this reasoning.

Keywords: deliberate reasoning, critical thinking, cognitive effort, heuristic response

Rationality: Can it be predicted by cognitive effort, ability and thinking disposition?

'Susans' parents have three children, April, May and..?'

This well-known children's riddle might not be the hardest to crack, however it's implications and more advanced 'siblings' receive a great deal of attention within psychology, economics, and in society as a whole. The use of gut-, and intuitive responses in leu of deliberate reasoning and critical thinking can lead to incorrect judgments and decisions not only in riddles and quizzes, but in everyday life, making us as some have said: predictably irrational (Ariely, 2008).

While it's hard to estimate the exact number of *meaningful* decisions and judgements a person makes every day, it's clear that in an ever increasingly complex world where we are bombarded with information to be evaluated and decisions to be made, making rational judgments and exerting critical thinking is of the upmost importance. Perhaps unsurprisingly then, 'critical thinking' is frequently rated as one of the top 'soft skills' managers want in their employees, and according to one US survey of managers this was rated at the very top, above communication, creativity and innovation skills (AMA, 2012).

This 'predictable irrationality' also has plenty of implications outside of the workplace. It has been linked with behaviour as diverse as who we vote for (Lau & Redlawsk, 2001), whether or not we overeat (Wansink & Sobal, 2007) and belief in the paranormal (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012). There have been written books upon books of all the large and small areas of life in which (ir)rationality has a huge impact (e.g., Kahnemann, 2011; Thaler, 2015; Thaler & Sunstein 2008; Ariely, 2008; Pennycook, Fugelsang, & Koehler, 2015).

Interestingly this rationality seems only moderately related to traditional measures of intelligence (for an overview, see: Stanovich & West, 2014), leading some researchers to call out for the need of a new measurement, a *Rationality Quotient*, seperat of the IQ-measures. One such attempt has been made in 'The Comprehensive Assessment of Rational Thinking' in the book 'The Rationality Quotient' (Stanovich, West, & Toplak, 2016) in which they argue that while common definitions of intelligence often allude to the concept of rationality, no current tests of intelligence actually incorporates and measures these aspects directly.

Interest in this gap between intuitive responses and normative correct responses, our 'irrationality', has spawned a whole sub-field in psychology: judgment- and decision making, or perhaps better known as the 'heuristics and biases'-literature. And it has made a bridge between economics and psychology in a field commonly referred to as 'behavioural economics' or 'neuroeconomics', seeking to bring psychological advances into models of human economic behavior (the two terms being used mostly interchangeably, although some differences exits, see Thaler, 2015).

In the center of this research area are the questions: *Why and when do we rely on these intuitive responses instead of engaging in deliberate reasoning?* To explore this and the mechanisms behind it, I will first review some of the current literature on human higher cognition, and then test some of the predictions these theories and models implicate.

Key Terms

Rationality, critical thinking, deliberate reasoning, and other related terms are often being used somewhat interchangeably within this literature. Sometimes referring to the exact same concept, other times referring to slightly different concepts, and sometimes referring to different concepts even when the same term is being used. Due to this, a short clarification is in order.

Our use of the terms are built on Stanovich et al. (2016)'s usage, which draws heavily from previous usage within behavioural science. Deliberate reasoning refers to the effortful process of thinking something through: actively deliberating and evaluating the different options in order to make your response. Critical thinking is very similar to deliberate reasoning and refers to the act or skill of deliberate reasoning, and sometimes partly your propensity to engage in it. Rationality are used in two ways, mostly in referring to the gap, or rather lack there of, between a normative correct, or logically derived response and any other response. The other broader way it can be used is in reference to whether you *should* do/ think/act in a certain way or not, however as this implies a judgment of normative or morally correct behaviour, this won't be used here[1].

System 1 and System 2

Within the judgment- and decision making field, our 'modes of operation' or ways of thinking, are usually divided into two distinct types of processing. This deviation of human higher cognition has a long tradition in psychology, and likely best known through the work of Kahneman and Tversky (e.g., Kahneman, 2011; Thaler, 2015). Their observations of systematic differences between normative correct responses and people's actual responses in different situations (e.g., Tversky & Kahneman, 1974) contributed to the development of a dual-process theory. Separating between our intuitive reasoning, system 1, and our deliberate reasoning, system 2 (eg., Stanovich & West, 2000; Kahneman, 2011; Evans, 1984).

System 1 is intuitive, automatic, unconscious, fast and effortless, and system 2 is deliberate, serial, conscious, slow and effortful (e.g., Kahneman, 2011; Stanovich & West, 2000). System 1's reliance on mental short-cuts, called heuristics, makes it 'fast and frugal' (Goldstein & Gigerenzer, 2002) however it can lead to predictably irrational responses (Ariely, 2008). And while system 2's reliance on deliberate reasoning often improves precision it comes at the cost of effort. This cost of effort is usually considered the main explanation for our tendency to prefer system 1's heuristics (although there are some disagreement, often dubbed 'The Great Rationality Debate', see Goldstein & Gigerenzer, 2002).

Since the original conceptualization of the system 1 and system 2 theory researchers have moved away from thinking of this as *one* duel-process theory with a whole set of defining features, and rather moved towards multiple 'duel-process dichotomies' for each of the defining features (Pennycook, De Neys, Evans, Stanovich, & Thompson, 2018). However these typically correlate and in most instances keep the broad strokes of system 1 and system 2 intact, making this a useful interpretation model of common clusters of duel-processes (for more on this, see Melnikoff and Bargh, 2018, and Pennycook et al., 2018's response, and Melinikoff and Bargh, 2018's counter-response). In particular the distinction between system 1 as a rapid and effortless process, and system 2 as a higher order deliberate reasoning process loading heavily on working memory, is typically kept (Evans and Stanovich, 2013). Making the system 1 and system 2 terminology useful for our purpose of investigating the differences in intuitive and deliberate reasoning.

Willingness to exert Cogntive Effort

System 2's cost of cognitive effort is usually considered aversive (Kool, McGuire, Rosen, & Botvinick, 2010), and when evaluating a course of action and potential rewards we tend to satisfice rather than optimize (Simon, 1955). This notion is at the core of much of the dual-process literature, however is mostly used as an underlying assumption or explanation, and rarely experimentally tested in itself (Kool et al., 2010).

This aversion to cognitive effort is an idea nearly a century old. Perhaps best exemplified by the seminal work *Principles of Behavior* by Hull (1943), in which he stated "other things equal, organisms receiving the same reinforcement following two responses which require different energy expenditures will, as practice continues, gradually come to choose the less laborious response." (Hull, 1943, p. 392). And while Hull was mainly discussing this principle in reference to laboratory rats traversing a labyrinth, this law was quickly used in reference to both humans and human cognition. Exemplified in Allport (1954; as pointed out in Kool et al., 2010)'s explanation of prejudice as due to humans' tendency to overgeneralize and categorize quickly as we don't like to exert effort.

The challenge in testing this 'law of less cognitive work' is that it's hard to separate out all non-effort based reasons for preferring the low effort path. A lower effort option might be preferred to minimize time on task, improve accuracy or maximize goal achievement (Kool et al., 2010), or due to differences in intellectual ability which might lead to differences in both perceived and actual effort demand (Kool et al., 2010; Westbrook, Kester, & Braver, 2013).

Two different approaches have been taken in order to experimentally investigate our supposed aversion to cognitive effort. The first by Kool et al. (2010) who created an implicit measurement of intrinsic motivation to exert effort. In this measurement a participant have to repeatedly choose between one of two cues and are then presented with one of two very easy tasks to solve ('is it an odd or even number' or 'is the number higher or lower than five'; see Figure 1), without any indication of their performance at the task. The hidden manipulation is that one of the cues repeats the former task 90% of the time while the other cue switches between the two tasks 90% of the time, thereby demanding more use of one's executive flexibility and thus imposing a higher effort demand. Kool et al. (2010) found support for an overall avoidance of cognitive demand (i.e. preference for the low switching cue) however this avoidance varied across participants. And according to Kool et al. (2010) this was partly due to individual differences in executive flexibility, as this might have affected the experienced effort demand difference between the two cues.

The second approach by Westbrook et al. (2013) who created a modified version of the N-back working memory task, in which willingness to exert cognitive effort is measured by an effort/reward-threshold. In this task participants undergo a normal N-back phase to establish their individual performance at the different levels. This individual performance is then their required performance in the experimental part, in which they have to explicitly choose between an easy 1-back task or a much more demanding N-back task (see Figure 1). A participant's willingness to exert effort for reward can then be measured by varying the amounts offered and observe the individual effort/reward-threshold. Westbrook et al. (2013) found an overall aversion towards effort, that increased with increased effort demands. However participants with high executive function showed less of an aversion and less of a effort level effect.

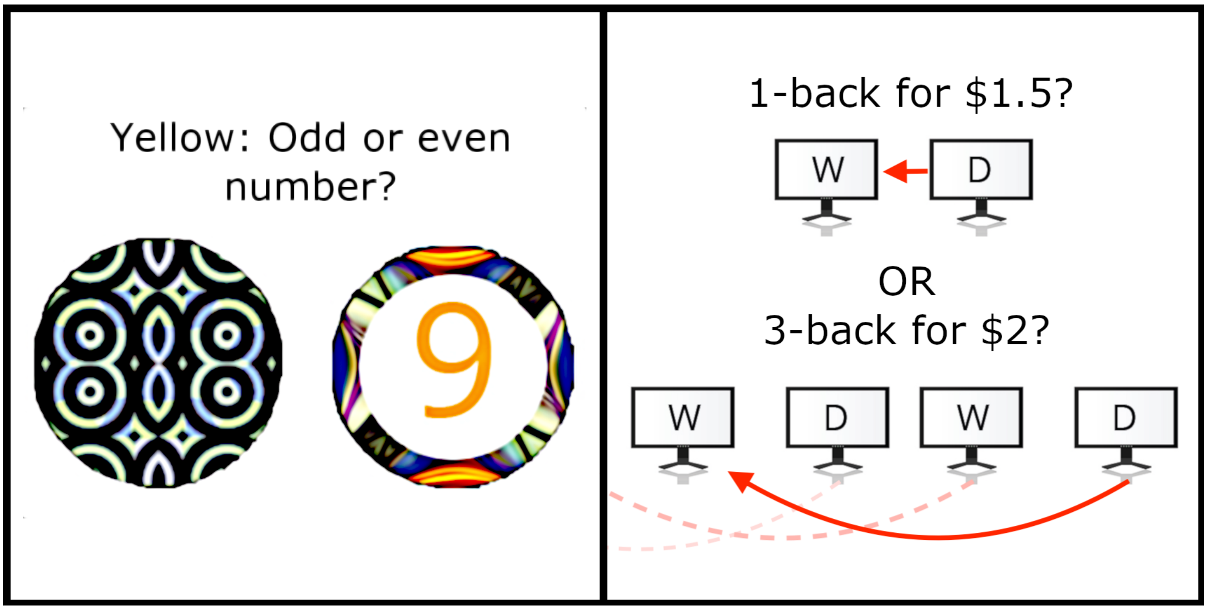


Figure 1. Left: Example of a trial in Kool et al. (2010)'s paradigm, after picking a cue, with the presented stimulus inside the cue (text not included in the actual task). Right: Conceptual example of a choice-screen in Westbrook et al. (2013)'s paradigm (task instructions not included in the actual task).

Taken together these two approaches provide support in favor of an overall aversiveness of cognitive effort (Kool et al., 2010; Westbrook et al., 2013), shedding some light on the *why* in our question: 'why do we rely on these intuitive responses instead of engaging in deliberate reasoning?'.

Individual Differences in Critical Thinking

The model of system 1 and system 2 postulates that one of the main reasons for *human's* reliance on system 1's heuristics and intuitive responses is due to the effort demands associated with system 2's deliberate reasoning. And while this was largely a hypothesis not yet scrutinized, newer research supports this aversiveness toward exerting cognitive effort (Kool et al., 2010; Westbrook et al., 2013). However we don't use system 1 all the time and even though the use of system 1 can be predicted in certain contexts (e.g., Kahneman & Tversky, 1979) it's mostly on a group level. With individual variability preventing predictions on when *individuals* engages in system 1 or system 2 reasoning, limiting the usefulness of this duel process model (Pennycook, Fugelsang, & Koehler, 2015).

To answer the *when* in our question we need to look at individual differences in the use of system 1 and system 2. As effort aversion seems to be one of the primary reasons behind system 1 thinking a willingness to exert effort should thus be a good predictor of deliberate reasoning. And while research on effort do find that the willingness to exert effort is influenced by cognitive ability, it is not wholly explained by it (Kool et al., 2010; Westbrook et al., 2013). Another factor that influence this willingness to exert effort might be an individuals thinking disposition. Thinking disposition is the individual's the propensity to engage in, or enjoyment of cognitively effortful processes, and one way of measuring this is by using the highly influential self-report measurement 'Need for Cognition' (NfC). Developed by Cacioppo and Petty in 1982, and quickly linked to individual differences in the use of system 1 versus system 2 reasoning (e.g., Cacioppo, Petty, Kao, & Rodriguez, 1986). And while highly useful (e.g., Fleischhauer et al., 2009) and demonstrated association with 'typical' system 2 reasoning (Cacioppo et al., 1986), it doesn't directly relay on the component which it's suppose to measure.

To address then 'when' in our question we need a tool to measure critical thinking more objectively than the NfC does. One such measure of system 2's deliberate reasoning is the Cognitive Reflection Test (Frederick, 2005; CRT). This simple three questions test measures an individual's propensity or ability to detect and resist a highly influential and available system 1 response, in order to provide a deliberate correct answer. Scores on this measure have shown a whole range of correlations with rational choices and a lower propensity to fall in other various 'heuristics traps' (e.g, Toplak, West, & Stanovich, 2011) even after controlling for cognitive ability (Toplak et al., 2011), and a stronger correlation with real-life measures such as SAT-scores than the NfC (Frederick, 2005). One possible explanation for this is that whereas the NfC measure an individual's tendency to, or enjoyment of exerting cognitive effort, the CRT measures an individuals cognitive reflection ability (Frederick, 2005), and this might be a better predictor of overall critical thinking and thus actual rational choices.

Linking these concepts: willingness to exert effort, thinking disposition, cognitive ability and (successful) cognitive reflection, within the system 1 vs. system 2 framework was the next logical step (e.g., Stanovich, West, & Toplak, 2016; Pennycook et al., 2015). To recapitulate: System 1 is fast, intuitive and automatic processing, if deliberate reasoning is needed, either through judged high importance of the outcome or through a detection of a conflict between our intuitive answer and the correct answer, cognitive effort is needed, i.e. activation of system 2. However, the amount of effort needed is determined by the task at hand, and mediated by one's cognitive ability (given that the task at hand benefits from increase cognitive ability). And the amount of effort we are willing to spend, is further determined by our propensity to engage in cognitive demanding thinking, and not least the judged importance of the outcome.

To incorporate those facets, Stanovich developed the tripartite model (Stanovich, 2009; Stanovich et al., 2016). It separates between the automatic mind, the algorithmic mind and the reflective mind. The automatic mind is very similar to system 1 from duel-process models, and this fast, intuitive and pre-attentive mind might show few if any, individual differences (Stanovich et al., 2016). In other words, while certain situations will influence the automatic response, this response will be very similar across people. Let us illustrate it with the now famous bat and ball problem: "A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?" (Frederick, 2005, p. 26). The system 1 or automatic response to this question, is always 10 cents, neither 1, 20 or 50 cents. The correct response is 5 cents, and would indicate deliberate reasoning in (naïve) participants.

The algorithmic mind on the other hand, refers to our maximum cognitive capacity, our potential to carry out deliberate reasoning. And according to Stanovich this is what intelligence tests (aims to[2]) measure, especially those assessing fluid intelligence and one's executive functions and working memory capacity (Stanovich et al., 2016; Stanovich, 2009).

The reflective mind refers to our rational thinking disposition, both through a willingness to exert cognitive effort, to use our algorithmic mind, and to our 'higher-order' thinking including concepts like goal directed behavior and critical thinking skill. One example could be a logic or math test, in which your algorithmic mind (mostly) determines the level of effort needed, and your reflective mind would determine whether or not this amount of effort is acceptable, given the extrinsic reward and your intrinsic motivation and disposition to exert the needed effort. Another example could be when judging if vaccines causes autism. You don't have any heuristic response, your algorithmic mind is capable to judge the evidence, and your incentive is high as you care for your child. However, if your critical thinking skill fails or your rational disposition isn't suitable enough, leading you to seek out wrongful-, or misjudge available information, you might end up with the wrong conclusion regardless. The reflective mind thus both initiates and determines the effort of the algorithmic mind, as well as act like a 'goal director' for it.

Stanovich and colleagues (e.g., Stanovich, 2016; Stanovich, West, & Toplak, 2016) used this tripartite model to lay out a prototype for a Rationality Quotient test[3] (for an overview, see Stanovich, 2016, table 3). And while some of it's items and subtests are dependent on knowledge (or made significantly easier with knowledge) it mostly draws on the concept of critical thinking, which is a prerequisite for or 'subspecies' of rationality (Stanovich, 2016).

To summarize, whereas the autonomous mind, or system 1, is fast and frugal, it can lead us astray. Critical thinking (successful deliberate reasoning) on the other hand requires both a well functioning algorithmic mind, our intellectual ability, and an attentive reflective mind, detecting and suppressing intuitive responses from the autonomous (system 1) mind, and controlling how and when to exert effort through our algorithmic mind.

Aims

Let's once again return to our question: 'Why and when do we rely on these intuitive responses instead of engaging in deliberate reasoning?' For the 'why', we have seen that effort seems to play a particularly large role. There are three different measures of an individuals willingness to exert cognitive demand: a) self-reported tendency, or enjoyment of demanding reasoning in the 'Need for Cognition'-scale (Cacioppo & Petty, 1982), b) the neuro-economic paradigm from Westbrook et al. (2013), and c) the implicit and intrinsic measurement from Kool et al. (2010). And while the NfC has been tested a lot against a range of other measures of rationality, e.g., CRT (Frederick, 2005; Toplak, West, & Stanovich, 2011), the two others haven't. In order to explore the influence of effort, we first need to establish a reliable measurement of effort. These three aforementioned tasks should be well related to each other if they measure the same construct. The paradigm from Westbrook et al. (2013) also supplies a good approximation of one's executive function/working memory (e.g., Haatveit et al., 2010; Oberauer, 2005). Whereas the paradigm from Kool et al. (2010) requires minimal working memory, but is more implicit in it's measure of willingness to exert cognitive demand than the explicit statements in the NfC.

For the 'when', if rationality depends on effort, cognitive ability, conflict detection, thinking disposition and knowledge, we can expect participants' performance on these three tasks above to be related to a measure of their critical thinking. This is commonly measured with the CRT, and this 3-item version has been extended to a 7-item version (see the methods section). All of these have a highly available intuitive, or heuristic response, requiring a detection and suppression of this response, in addition to successful deliberate reasoning in order to reach the correct answer. We also included some more items from the literature (see methods section) drawing inspiration from the 'non-heuristic' subtest's included in Stanovich (2016)'s Rationality Quotient. This way we can separate the role of intuitive responses, from either a lack of motivation or algorithmic ability.

And lastly, we can then ask: which of the three tasks assessing cognitive effort best predicts the performance in the critical thinking task? In more detail, will the Kool et al. (2010) paradigm which uses only intrinsic motivation and sensitivity to smaller changes in effort be a better predictor than the neuro-economic and 'algorithmic'-heavy paradigm from

Westbrook et al. (2013)? Or do people have a good calibrated reflective mind and the NfC works just as well as the experimental approaches?

Specifically we had four hypotheses, the first three to assess our measurement of willingness to exert cognitive effort: a) Will the measurements from Kool et al. (2010) and Westbrook et al. (2013) capture the same concept? b) Will the measurements from Kool et al. (2010) and NfC (Cacioppo & Petty, 1982) capture the same concept? and c) To examine our effort measurement, how stable is this measurement of willingness to exert effort, i.e., does it show good test-retest reliability? And the last hypothesis to examine the role of willingness to exert cognitive effort in critical thinking: What is the relationship between one's score on the extended critical thinking task (our 'Rationality Quotient'-task) and willingness to exert cognitive effort?

Methods

We used an observational within-subject design with eight different computerized tasks and measurements, of which four will be discussed in this paper. The tasks measure the propensity to engage in deliberate reasoning and critical thinking, intellectual ability, and different aspects of willingness to exert cognitive effort.

The tasks were administered to a diverse group of (semi-)blinded participants. The participants were tested individually, in two sessions, with one to two months between sessions. Using the same non-blinded experimenter in all sessions, but with a limited amount of oral instructions, and with a strong adherence to an experimenter-script where applicable.

Preregistration and Open Science

This project was pre-registered on the Open Science Framework (OSF), for both the Collaborative Replications and Education Project replication part and for the overall project (https://osf.io/2zw3v/; https://osf.io/yheqd/).

As per Nosek, Ebersole, DeHaven, and Mellor (2018; see also Grahe, 2014)'s recommendation, a strong distinction between pre-registered prediction testing, and exploratory testing were drawn. To do this we separated the pre-registred confirmatory analyses, which can be found on the pre-registration form on OSF (https://osf.io/yheqd/), from the exploratory analyses in all subsequent sections. In addition we separated our secondary hypotheses from the post-hoc explorations. This does not imply that the predictions and explorations were generated to 'fit the data', but rather that the data prompted new interesting predictions. All non-confirmatory hypotheses, both secondary hypotheses, and post-hoc explorations must be regarded as exploratory and hypotheses-generating, not confirmatory.

In order to facilitate future replication (e.g., Munafò et al., 2017) and meta-analytic efforts, all raw data, analyses, and materials used were uploaded to OSF (https://osf.io/yheqd/; page will be opened upon article publication, please contact the authors if you wish access prior to this). With one exception: for the Demand Selection Task Debriefs the answers were provided by hand-writing and only the experimenter's interpretation of these answers were uploaded, in order to keep our participants anonymity intact.

Participants

Forty participants (aged 18-35; 27 women and 13 men) accepted the invitation and participated. The participants had been told they would receive a fixed non-monetary reward for participating, but were in addition given a small monetary reward, ranging from 50 NOK to a maximum of 150 NOK (approximately 16 USD) based on summary performance on two tasks. The performance dependent reward applied for one task in session one, the physical effort measuring EEfRT task, and one task in session two, the cognitive effort measuring COG-ED task. The participants were told about this extra reward opportunity in the beginning of the first reward-earning task, and that this *only* concerned their performance on these two tasks.

The participants were (semi-)blinded in that we had only told them it was a psychological study within the field of cognitive psychology. In addition we started with the least explicit task, before moving to the more revealing tasks. The participants were not told about the goals or hypotheses of the overall project, nor for the individual tasks, prior to completion of session two.

Sample size estimation and stopping criterion. Our sample size was based on an expected Spearman's rank correlation coefficient of at least .50 across four different hypotheses. With an accepted type I error of 5%, and an accepted type II error of 20%. Using a Bonferroni-correction for running four different tests (e.g., Miles & Field, 2010).

Using G*Power 3.1 (Faul, Erdfelder, Lang & Buchner, 2007; Faul, Erdfelder, Buchner & Lang, 2009) an N of at least 41 was recommended.[4] As this was part of a master's thesis, with time- and expenditure limitations, we sat the stopping criterion at the suggested N.

Inclusion. The following criteria had to be met in order to participate: between and including, 18 to 50 years old, with normal- or corrected to normal eyesight, no psychiatrical or neurological disorder, no drug use within three months prior to the testing sessions (excluding tobacco, caffeine, nicotine, and alcohol, although participants were encouraged to not 'binge'-drink the day prior to testing) and no current intake of central nervous-system medications (e.g., anti-depressants, anti-epileptic drugs, or ADHD-medications like methylphenidate, Ritalin and Concerta).

A signed informed consent form, including these criteria (see Appendix A for the informed consent form) were required prior to any participation.[5]

Participants had to be fluent in Norwegian as we wanted to limit any potential bias by providing the task instructions in multiple languages. In addition participants had to indicate that they would participate in test session two, four to eight weeks after session one.

We also wanted to limit the overall number of psychology-students participating, as they might have greater experience-, or knowledge of the instruments used, preventing a potential greater chance of discovering what the project's aims were. This limit was set to 50% and all participants had to indicate if they were psychology-students or not.

Recruitment. The participants were mainly recruited via e-mail. An invitational e-mail was sent to the study-advisors at the 30 largest study programs at UiT The Arctic University of Norway (UiT), including all different faculties of this broad-spectrum university (excluding psychology-programs). This in order to get a representative sample, and avoid sampling bias as much as possible. This effort yielded a total of 34 participants.

Potential participants were told that they could forward the invitation to others, and an additional six participants were recruited through this convenience sampling, of which three were full-time workers and three were high school students (aged 18 or above).

The invitational e-mail[6] was as vague as ethically permissible to avoid recruiting participants that were especially fond of puzzles, brain-teasers, et cetera, that could bias our results.

Ethics. The project was evaluated and approved by the institutional review board at the Department of Psychology, UiT (see Appendix B for the ethics application).

Participants read and signed the informed consent form prior to participating (see Appendix A for the informed consent form). They were encouraged to ask any questions they might have regarding the consent form, and were given a brief summary of the most important aspects of the informed consent form: their right to full anonymity, insight into their contributed raw-data, and the right to withdraw their consent and participation at any time without providing a reason.

The participants were given a short debrief of the overall aim's of the project, as well as what the different tasks measured following session two. All participants were invited to a more extensive, collective debrief session, in which they could get their raw-scores across all tasks if they desired. Anonymity were kept by using an electronic sign-up sheet were participants signed up for the debrief session using only their ID. Envelopes with the individual scores were marked with the ID's, and participants picked them up themselves in the beginning of the debrief session.

The distinction between individual prediction versus group-wise predictions were stressed. Both in the short debrief following task completion on day 2, as well as in the extensive debrief session. The limitations in the project and tasks were explicitly mentioned.

All participants earned something in the 'extra reward'-tasks, the extra rewards ranged from 50 NOK to 150 NOK.

Location and Site

The research was conducted in a psychology-lab at UiT Campus Tromsø, Norway. The participants were tested individually in both sessions, in a small noise-isolated computer room without any distracting elements. The experimenter left the room prior to all tasks unless otherwise noted, and was notified by the participants upon completion of the different tasks.

Materials and Procedures

The participants were tested in two sessions, with the tasks being administered in the order as presented in this section, see Figure 2, and Figure 3 for an overview of the testing sessions. The tasks with a dotted-line were administered, but are not discussed in this thesis.

They will be briefly described as they might have affected the participants' performance and responses in the included tasks and measurements, but their results will not be presented nor discussed.

All instructions were given in Norwegian.[7]

Session 1.

Demand Selection Task (DST). This task was developed by Kool et al. (2010), and is a computerized task that is meant to implicitly measure intrinsic willingness to exert cognitive effort. Implicitly in that it doesn't tell the participants that there's a difference in cognitive effort demand associated with the different cues (the manipulation). Intrinsic in that the participants aren't given any rewards, scores or otherwise extrinsically driven motivation to choose one of the cues (demand level) above the other. A participant's preference in favour of the low-demand cue to the high-demand cue, is taken as a measure of their aversion towards expending cognitive effort.

This specific task was in addition a part of an international replication project, through the Collaborative Replications and Education Project (see https://osf.io/2zw3v/ for the full preregistration, including all material needed to fully replicate). Replicating Kool et al. (2010)'s Experiment 3, with the ergonomic and bias-reducing changes introduced in Experiment 5 (but without the preliminary block used to calculate 'switch cost').

The task was administered on a computer, using MatLab 2018a (The MathWorks, MATLAB, Version 9.4, 2018), with the Psychophysics Toolbox 3 extension (Brainard, 1997; Pelli, 1997; Kleiner, Brainard, & Pelli, 2007). Only minor technical and non-significant changes had to be done to the original task script (see https://osf.io/2zw3v/ for a full explanation of the script changes).

Prior to starting the recorded part of the task, the participants went through a training-session. In the training-session participants got instructions on how to respond to the different stimuli, and were provided with a hand-out-script of the instructions (see Appendix C for the experimenter-script; https://osf.io/2zw3v/ for a recorded pilot-run of this task with subtitles; see Appendix D for both the used hand-out, and the English translation of the hand-out). Participants could refer to this hand-out should they forget the instructions. Participants were then sequentially presented with stimuli, without any cue selection, and they got instant feedback upon responding, through either a green or red dot, for the first 20 training-trials, and for the subsequent 40 training-trials they got summary feedback after each 10th trial. Upon completing all 60 training-trials with sufficiently high scores (no participant scored below 56 out of 60 in these training-sessions), the participants gave notice to the experimenter and as a final part of the training-session participants got four training-trials with cue selection (see Figure 1).

After completing the training-session, participants were told that the actual task would begin, and how to proceed in that task. The participants were instructed to do the same as they had done in the training-session, but with the addition of choosing cues in order to be presented with stimuli. They were also told there was no time limit, and that they should try out both cues, not by using 'simple rules' (e.g., alternating) but rather that it should feel like they were making a decision for each trial however if they should start to favour one cue, they could choose that cue as much as they wanted.

In each trial participants had to choose a cue, and were then presented with a stimulus, of which a response had to be made, before moving on to the next trial. Participants chose between one of two colourful circles on the screen, the cues, by moving the mouse cursor to their selected cue, and were then presented with the stimulus inside of their selected cue (see

Figure 1). The stimulus was a single-digit Arabic numeral, between and including, one and nine (with the exception of the number five), in either yellow or blue colour. The correct response to the stimulus depended on both the colour of the numeral and of the numeral itself. When the stimulus was a blue numeral, participants had to make a magnitude judgement, clicking the left-hand side mouse button if the numeral was below five, or the right-hand side mouse button if the numeral was above five. When the stimulus was a yellow numeral, participants had to make a parity judgement, clicking the left-hand side mouse button if the numeral was an odd number, or the right-hand side mouse button if the numeral was an even number. After responding to the stimulus, the cue went back to normal, although now with both cues appearing in a dimmer light, indicating that the trial was over. Participants had to move the mouse cursor to a small white dot located exactly in the middle of the space between the two cues, in order to "re-activate" the cues (i.e. make the cues bright again) in order to minimize any cue preferences due to ease of hand-movement. The next trial was then ready to start, and participants proceeded by picking a cue again.

Unbeknownst to the participants, the two cues differed in their stimulus-response task-switching rate. In every trial, one cue had a task-switching rate of 0.1 (the low-demand cue), and the other had a task-switching rate of 0.9 (the high-demand cue). The low-demand cue thus had a 90% chance of presenting participants with the same stimulus-response task (i.e. with the same colour of the numeral) as in the trial preceding it. While the high-demand cue had a 90% chance of presenting participants with the opposite stimulus-response task (i.e. with a switched colour of the numeral) as in the trial preceding it. Within each block the cues' appearance and location stayed the same, as did the cues' stimulus-response task-switching rate. Between blocks the cues' appearances and locations changed, both in regards to exact screen location, and the relative position between the cues (e.g. going from one of the cues being above the other cue, to both of the cues being on a line, with one cue to the left-hand side and the other cue to the right-hand side). Most importantly this meant that if a participant had found the task-switching manipulation in one block, either consciously or unconsciously, the participant would've had to search for it, or rediscover it again in the next block, if they preferred to stay on one specific demand-level. This also decreased the possibility of specific demand-level being preferred by accident (e.g. because it always were on the cue to the left-hand side, or on the "prettiest" cue). The individual participant's overall selection of the low stimulus-response task-switching cue against the high stimulus-response task-switching cue was the crucial measurement we wanted to make with this task.

The participants underwent a total of 600 trials, divided into eight blocks with 75 trials in each block.

Following the DST task a paper-and-pencil debrief questionnaire were administered to the participants. The debrief asked participants open-ended questions on what it was like performing the task, how they chose between the circles, and whether or not they felt like they developed a preference for one circle (cue) to the other (item 1 - 3, see Appendix E for both the used debrief questionnaire, and the English translation of the debrief questionnaire). This was done in order to try to catch any manipulation discovery, without increasing participants knowledge of the manipulation prior to the re-administration of the DST task in session 2. It is important to note that this was a deviation from Kool et al. (2010)'s Experiment 3, as they administered the full debrief questionnaire following the task-session. The full debrief questionnaire explicitly tells the participants about the manipulation, in order to ask them whether it seemed like this manipulation was present in their task. However doing so in our project could potentially decrease the validity of the test-retest, so we postponed this to session two.

Rationality Quotient (RQ). This task consisted of 14 items from the judgment-, and decision-making literature. The items chosen are often argued to be a measure of deliberate reasoning and critical thinking (e.g., Frederick, 2005; West, Toplak & Stanovich, 2008; Toplak, & Stanovich, 2002) and subsequently a part of-, or a prerequisite for rationality (e.g., Stanovich, West, & Toplak, 2016; see also Stanovich, 2016).

The items can be divided into two sub-categories, items *with* an incorrect heuristic response, and items *without* a heuristic response (see Appendix F for an overview of all items, both as given and the English translation). All of the items required successful deliberate reasoning in order to find the correct answer, however the heuristic items involved a detection and suppression of the incorrect heuristic response as well.

The task was administered on a computer, through Qualtrics (Qualtrics, Provo, UT), and the participants were presented with the items in a mixed order from the sub-categories, one item at the time, in the same order for all participants. Some of the items had specific answer alternatives, while others had open-answer fields (see Appendix F). No time-limitation, nor time-tracking were indicated to the participants, they were only told that for the next task, they would have to solve some exercises (see Appendix F for the written intro given in this task).

The heuristic sub-category consisted of seven items, of which six items were Cognitive Reflection Test items from Toplak, West, and Stanovich (2014, p. 151, CRT7; adopted from Frederick, 2005, p. 27, CRT1-3; personal correspondence between Toplak, West, & Stanovich with Frederick, 2011, CRT4-5; adapted from Dominowski, 1994, CRT6). In addition, we had one probability matching item (Koehler & James, 2010, p. 669).

An example of a heuristic item, would be item 4, "It takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?" (Frederick, 2005, p. 27, CRT2). The number '100' pops-out as an intuitive answer, but 5 is correct.

The non-heuristic sub-category consisted of seven items of which five items were without a heuristic answer, and two items had a possible, but not definite heuristic answer. Of the five items without a heuristic answer, three items were dependent on Bayesian reasoning. One probability estimation item (Teigen & Keren, 2007, p. 339), one conditional probability item (G. Gigerenzer, 2007; as cited in Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007, p. 55) and one prior-posterior Bayesian item (Stanovich, West, & Toplak, 2016, p. 100; built on Stanovich & West, 1998; adapted from Beyth-Marom & Fischoff, 1983). The two other non-heuristic items were one conditional reasoning item (Lehman, Lampert, & Nisbett, 1988, p. 442; similar to Wason, 1966) and one covariation detection item (Toplak, West, & Stanovich, 2011, p. 1285).

The two items with a possible but not definite heuristic answer, were both boolean/binary disjunctive reasoning items (Levesque, 1986, p. 85; Smullyan, 1978, p. 22 as cited in Toplak & Stanovich, 2002, p. 201; Rips, 1989; see Appendix F, item 2 & 10).

An example of a non-heuristic item, would be item 10:

> Imagine that there are three inhabitants of a fictitious country, A, B, and C, each of whom is either a knight or a knave. Knights always tell the truth. Knaves always lie. Two people are said to be of the same type if they are both knights or both knaves. A and B make the following statements: A: "B is a knave!" B: "A and C are of the same type!" What is C? A knight, a knave, or cannot be determined? (Smullyan, 1978, p. 22 as cited in Toplak & Stanovich, 2002, p. 201; Rips, 1989)

In this item a heuristic response of 'cannot be determined' might appear, but regardless of this, in order to arrive at the correct solution one has to do some boolean logic deliberation, leading to 'knave' as the correct answer.

Following these 14 items, participants were given a debrief question, in which they indicated how many of these items they had encountered prior to this task.

BullShit Receptivity (BS). The task is from Pennycook, Cheyne, Barr, Koehler, and Fugelsang (2015) and shall measure an individual's receptivity to pseudo-profound bullshit. Participants' conflict detection are separated from the participants' general reflective thinking propensity, by comparing the participants' ratings for pseudo-profound statements to the their ratings for motivational quotations.

The task was administered on a computer through Qualtrics (Qualtrics, Provo, UT). It immediately followed the previous task within the same Qualtrics-form. The task consisted of a mix of 10 pseudo-profound statements and 10 motivational statements (Table S1 and Table S5 in Pennycook et al., 2015).

The participants were to indicate how deep of a meaning each statement had, on a scale from 1, not deep meaning at all, to 4, very deep meaning.

This task is not further discussed in this thesis.

Need for Cognition (NfC). This self-report instrument was made by Cacioppo and Petty (1982; Cacioppo, Petty, & Kao, 1984, 18-item short version), and measures temporally-stable individual differences in one's tendency-, or likelihood of, enjoying, seeking, or engaging in intrinsic motivated effortful cognition (Cacioppo, Petty, Feinstein, & Jarvis, 1996). Temporally-stable, both as in the theoretical construct (e.g., Cacioppo et al., 1996), and later empirically supported for both shorter (Sadowski & Gulgoz, 1992) and longer periods of time (Bruinsma & Crutzen, 2018). Intrinsic in that it only measure an individual's motivation to engage in cognitive demanding tasks in absence of-, or with minimal extrinsic reward (Thompson, Chaiken & Hazlewood, 1993).

This self-report measure was administered in Qualtrics (Qualtrics, Provo, UT). It immediately followed the previous task within the same Qualtrics-form.

We used the 18-item short-version (Cacioppo et al., 1984), translated to Norwegian (see Appendix G, for both the translated version, and the original English version). For each of the 18 statements participants were to indicate how well each statement described them, on a scale from 1, very uncharacteristic of me, to 5, very characteristic of me. Of the 18 items, nine items were statements indicative of high 'need for cognition', and nine items were indicative of low 'need for cognition'.

An example of a 'high need for cognition'-item would be item 6, "I find satisfaction in deliberating hard and for long hours" (Cacioppo et al., 1984, p. 306). An example of a 'low need for cognition'-item would be item 16, "I feel relief rather than satisfaction after completing a task that required a lot of mental effort" (Cacioppo et al., 1984, p. 306).

Effort Expenditure for Rewards Task (EEfRT). This task was developed by Treadway, Buckholtz, Schwartzman, Lambert, and Zald (2009) measures an individual's reward motivation and effort-based decision making in tasks concerning physical effort.

The task was administered through Inquisit 5 Web (Inquisit, Millisecond Software, 2018).

During a fixed 10 minute task participants underwent several trials were they were to chose between an effortful task or less effortful task, with varying potential rewards. For the effortful task participants had to click 100 times on the keyboard spacebar with their pinky-finger on their non-dominant hand in less than 21 seconds. For the less effortful task, participants had to click 30 times on the keyboard spacebar with their thumb on their dominant hand in less than seven seconds. By varying the odds for reward and varying the reward sums, we could measure a participant's propensity to engage in physical effortful work. Participants were given extra reward in this task, ranging from 0 NOK to 50 NOK, paid out together with the other extra reward task (COG-ED) following session two. It was explicitly stated that this extra reward opportunity solely concerned their performance on these two tasks, and not influenced in any way by their performance in any other task.[8]

This task is not further discussed in this thesis.

NASA Task Load Index (N-TLX). The NASA Task Load Index (N-TLX) was developed by Hart and Staveland (1988) and is a self-report measurement of perceived workload, effort and self-rating of one's performance during other tasks. The N-TLX consists of six items on which participants are to rate their perceived mental effort needed on the task, perceived physical effort needed on the task, perceived temporal pressure in the task, self-reported performance satisfaction, perceived effort (mental and physical) invested into the task by the participant, and level of frustration felt during the task. Participants responds to each item, using a scale from 0, very low, to 100, very high.

The measurement was administered in Qualtrics (Qualtrics, Provo, UT). The N-TLX was administered two times: following the DST Debrief, before the RQ-task, and following the RQ-task, before the BS-task.

This measurement is not discussed further in this thesis.

Procedure session 1. The participants were greeted in a waiting area were they read and signed the informed consent form. They were encouraged to ask any questions regarding the form, should they have any, and then the experimenter briefly repeated the most important parts of the form.

Participants were given a three digit participation ID after signing the informed consent form, encouraged to write it down on their phone, turn the phone off, and then lead into the computer-lab. They were then presented with the tasks and measurements in the order as previously described (see Figure 2).

The session took approximately one and a half hour, and these sessions were held in November and December, 2018.

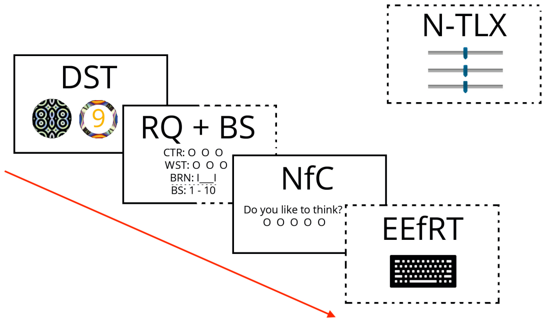


Figure 2. Overview of the task sequence in session one. The tasks with a dotted line were administered as well, but are not discussed in this thesis.

Session 2.

Demand Selection Task (DST). The task was re-administered to the participants exactly in the same way as described in session one. With the same experimenter-script being followed, and included the training-session.

The only deviance from the first session was that the participants were given the full debrief questionnaire, all six items, upon completion of the task (see Appendix E for both the used debrief questionnaire and the English translation of the debrief questionnaire).

Need for Cognition (NfC). The instrument was re-administered to the participants exactly in the same way as described in session one.

The only deviance from the first session was that this time it immediately followed the DST Debrief and DST N-TLX, as opposed to following the BS-task.

Handgrip effort task. This task measures an individual's intrinsic willingness to exert physical effort. Intrinsic in that the participants aren't given any external reward or extrinsically driven motivation to exert effort (they are however provided with a 'virtual reward').

The task was administered on a computer, using MatLab 2018a (The MathWorks, MATLAB, Version 9.4, 2018; see https://osf.io/yheqd/ for the task script) connected to a hand-dynamometer (Hand Dynamometer, HD-BTA, Vernier).

Unbeknownst to participants the task consisted of three rounds. In the first round participants were told to squeeze as hard as they could during a short 10 second trial (had to be able to hold that pressure for one second). In the second round, participants are told to squeeze hard enough for a black-dot to appear on the screen, and that their goal should be to keep that dot visible for as long as possible during a 60 second trial. This second round was repeated in a third round.

The participants max strength were recorded in round one, and without participants knowing, used to set the reference for when the black-dot appeared in round two and three (at 70% of max). Since the participants were thus 'competing' against themselves, their total time above the 70% reference-point was a measure of their willingness to exert physical effort.

This task is not further discussed in this thesis.

Cognitive Effort Discounting Paradigm (COG-ED). This task was developed by Westbrook et al. (2013), and is a computerized task that measures explicit willingness to exert cognitive effort through reward discounting. Participants repeatedly choose between an effortless task and an effortful task, with varying rewards, and their effort-reward threshold is taken as a measure of their willingness to exert cognitive effort. The first phase of this task is a N-back task (originally created by Kirchner, 1958) which measures the executive function of working memory (Owen, McMillan, Laird, & Bullmore, 2005).

The task was administered through Inquisit 5 Web (Inquisit, Millisecond Software, 2018) and used a shorter 1 to 4-back version of the original task (see https://osf.io/2zw3v for the script).

Prior to starting the recorded part of the task, the participants went through a training-session. In the training-session participants got instructions on how to respond in the different N-back levels and were then presented with a practice block of nine trials for each level (1, 2, 3, & 4 back).

The first phase of this task consisted of five runs per N-back level (2, 3, & 4), each run with 5 target trials (response would be correct), and 10+N non-target trials (response would be incorrect) in a pseudo-random sequence. Each trial lasted 2.5 s, and in each trial participants were presented with a stimuli (one of 20 consonants, centered white letters on a black screen, sans-serif font) for 0.5 s, followed by a black screen for 2.0 s, and during this trial time had to either respond (press 'A' on the keyboard) or not respond. The correct response depended on the current N-back level, if the presented stimuli were the same stimuli as N-trials previously, the correct response were to respond, if it was not the same stimuli as N-trials previously, the correct response were to not respond. After each run, the participants were presented with a summary feedback of their accuracy, and after the last run on each N-back level they were presented with a level summary.

The second phase consisted of three blocks, 1-back vs. 2-back, 1-back vs. 3-back, and 1-back vs. 4-back, presented in a pseudo-random order across participants. Each block had six runs in which the participants chose between a 1-back task or N-back task. The tasks themselves were equal to the N-back task described above.

In choosing the N-back task, the participants were given a fixed 2$ if their performance was as good as, or better than their performance on the specific N-back level in the first phase (this was explicitly told to the participants at the choice screen, together with their specific performance). In choosing the 1-back task, the participants were given an adjusted amount if their performance was above 80%. For the first round all participants were offered 1$ for choosing the 1-back. For the subsequent levels, each adjustment were half of that in the previous round, and was adjusted up if the participant chose the N-back task, and down if the participant chose the 1-back task. This adjustment was reset between each of the three blocks.

In example a participant in the 1-back vs. 3-back task block would be offered 2$ for choosing the 3-back and 1$ for choosing the 1-back in the first run. If the participant chose the 3-back, they would be offered 2$ for the 3-back and 1.5$ for the 1-back in the second run (see Figure 1). If the participant then chose the 1-back, they would be offered 2$ for the 3-back and 1.25$ for the 1-back in the third run, et cetera. Each time with half as large adjustment as in the previous run, until all runs within a block were completed and the adjustments were reset for the next block.

After each run the participants were told whether they got the reward or not, and after each block they were given a total earnings this far in the task. Upon completion of the task, the participants got 3x their earned amount in NOK, together with their earnings in the EEfRT from session 1.

NASA Task Load Index (N-TLX). The measurement was re-administered to the participants the same way as described in session one. The N-TLX was administered three times: following the DST Debrief, before the NfC-instrument, following the Handgrip Effort task, before the COG-ED task, and following the COG-ED task.

Procedure session 2. Upon completion of all test session ones a sign-up form for session two were sent out through e-mail to all participants. This e-mail encouraged them to choose a date within four to eight weeks of their first session, and to choose a time-slot roughly equal to their first time-slot in order to minimize any systematic differences in wakefulness and alertness in the two sessions. This was especially important and stressed for participants that chose to participate very early or very late in the day (see Appendix H for an overview of the session dates and times).

Participants was greeted in the same waiting area as in session one and lead into the computer-lab. They were then presented with the tasks and measurements in the order as previously described (see Figure 3).

The session took approximately one and a half hour, and these sessions were held in January, 2019.

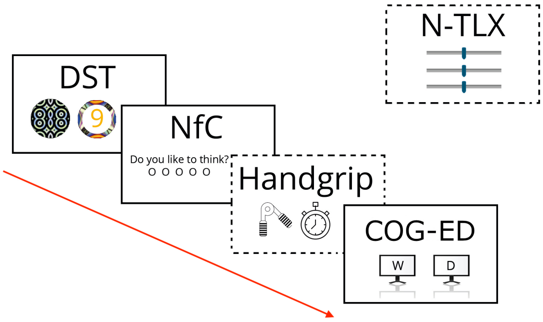


Figure 3. Overview of the task sequence in session two. The tasks with a dotted line were administered as well, but are not discussed in this thesis.

Data Collection and Analyses

In accordance with current directions in (psychological) science and recommendations from The American Statistician (e.g., Wasserstein, Schrim, & Lazar, 2019; Wasserstein & Lazar, 2016; see also Munafò et al., 2017; Nuzzo, 2014), *p* values will be disclosed but not commented upon nor denoted.

Data management. All collected data were only identifiable via a three digit ID, and these ID-numbers were never connected to the participants' names in any way.

The raw data from MatLab (DST1, DST2, and Handgrip effort task), Qualtrics (RQ, BS, NfC1, NfC2, and N-TLX) and Inquisit Web (EEfRT, and COG-ED) were uploaded to OSF, on a server located within the EU (Germany) and were thus protected by The EU General Data Protection Regulation.

Summary variables were created and organised in Microsoft Excel, and stored as CVS-files on OSF. Statistical analyses were carried out in JASP (JASP Team, 2019, version 0.9.2) and R/RStudio (R Core Team, 2018, Vienna, Austria; RStudio Team, 2016, Bosten, MA), using the 'Rfit'-package (Kloke & McKean, 2012) for ranked-based estimation of linear models.

Data collection and variable calculations.

Demand Selection Task (DST). All DST results from session one were labeled *DST1* and all DST results from session two were labeled *DST2*.

For each participant, in each trial, we recorded: the cue-selection (whether the cue was a high-demand or low-demand cue), the presented stimulus-task (whether the task was a repeated or a switched task), the response (whether the response was correct or incorrect), the response-time (measured from the presentation of the stimulus to the response was made), and the trial-number. The first trial in each block was disregarded, as this trial couldn't be regarded as neither a repeated- nor a switched-task trial.

The main measurement of interest in this task was a participant's low demand preference (*DST*-LDP*). This was calculated using the ratio of low-demand cue chosen to high-demand cue chosen. Ranging from 0, all high-demand cues chosen, to 1, all low-demand cues chosen. A low-demand preference of .50 would indicate no specific demand preference.

To detect if any participant had to be excluded we calculated the individual participant's accuracy (*DST*-ACC*), using the ratio of response-correct to response-incorrect. Ranging from 0, no correct responses, to 1, all correct responses. An accuracy of .50 would indicate random-clicking as every trial had a binary response with one correct and one incorrect response.

For the requirement that the high task-switching rate were indeed more cognitively demanding than the low task-switching rate we refer to Kool et al. (2010)'s Experiment 5. Any attempt to use our observed accuracy or response-time differences between the two demand-cues or task-switching rates would not be indicative of the actual effort demand differences. They might be a result of, or at least heavily affected by the very thing we wanted to measure, the low demand preference (for further discussion of this, see Wylie & Allport, 2000; Kiesel et al., 2010; Liefooghe, 2017). This does however concern the validity claim of the DST and will be revisited in the discussion.

For the debrief questionnaire the open-ended hand-written answers were interpreted and coded for three different aspects. The first, detected manipulation (*DST*-DM*), where 0 was no manipulation detected, 0.5 was a partial detection, and 1 was manipulation detected. The second, developed a preference based on technical aspects (*DST*-Tp,* e.g., ease of hand-movement or better visual contrast between the cue and the stimuli), where 0 was no technical preference noted, and 1 was technical preference noted. The third, developed an unrelated preference (*DST*-Up*, e.g., prettiest or coolest cue), where 0 was no other preference noted, and 1 was other preference noted.

Rationality Quotient (RQ). The main measurement of interest in this task was the total score across all 14 items. This score was labeled *RQ*, and went from 0, no item correctly answered, to 14, all items correctly answered, with all items having equal weighting.

To detect if any participant had to be excluded we used the debrief question(*RQ-Db*), in which the participants indicated their prior knowledge of the items used. 0 indicted no prior experience, 1 indicated experience with a few of the items, 2 indicated experience with almost half of the items, 4 indicated experience with more than half of the items, and 5 indicated experience with nearly all items. After the exclusion, the variable were recalculated into 0, no experience and 1, any experience.

Twelve of the items had one specific correct answer and this was coded 1, all other answers were coded 0. Two items (item 9 and item 13) were coded the same way, but with a wider range of what was considered correct. Item 9, the prior-posterior Bayesian item, consisted of two parts and this item was considered correct when the participants indicated a lower posterior probability in part two than the prior probability the individual participant provided in part one. Item 13, the covariation detection item, was considered correct when the participant provided an answer below 0. This was done in order to see if the participants understood the direction their answers should have, without requiring the exact calculations to be performed successfully.

For the post-hoc exploratory hypotheses we used the sub-categories. All of the non-heuristic items (item 2, 5, 7, 9, 10, 12 and 13) were scored as described above and the total score was labeled *RQ-nH*. All of the heuristic items (item 1, 3, 4, 6, 8, 11 and 14) were scored as described above and the total score was labeled *RQ-H*. Both went from 0, no item correctly answered, to 7, all items correctly answered. As a measure of a participant's heuristic response suppression the variable *RQ-HRS* was created, in which all correct or non-heuristic incorrect answers were coded 1, and all heuristic answers were coded 0 (see https://osf.io/yheqd/ for the raw data).

Need for Cognition (NfC). For each participant the summary NfC score from session one were labeled *NfC1* and the summary NfC score from session two were labeled *NfC2*.

A summary score was calculated by adding all of the items, with equal weighting. Item 1, 2, 6, 10, 11, 13, 14, 15 and 18 were 'high need for cognition'-items, and were added as they were provided. Item 3, 4, 5, 7, 8, 9, 12, 16, and 17 were 'low need for cognition'-items, and were reversed (6 minus item response) prior to summary into the total score. The total score went from, 18, very low 'need for cognition', to 90, very high 'need for cognition'.

Cognitive Effort Discounting Paradigm (COG-ED). This task consisted of two phases, for the first phase (the 'normal' N-back) the main measurement of interest was a participant's performance (*COG-ED d'*). This was calculated by averaging their signal detection *d'* in the 2-back, 3-back, and 4-back blocks in the first phase (excluding the first practice block). The signal detection was calculated as *d'* = Z(Hit) - Z(FA), where Hit = hit/(hits+misses), and FA = false alarms/(false alarms + correct negative). In the case of perfect scores, Hit was calculated as 1-1/(2n), and for zero false alarms, FA was calculated as 1/(2n), where n was the number of total hits or false alarms (Macmillan & Creelman, 1990; as cited in Haatveit et al., 2010). For the five blocks in each N-level (2, 3, & 4), there were 5 target trials, and 10+N non-target trials. Yielding a theoretical max score of 4.45 in the 2-back, 4.48 in the 3-back and 4,50 in the 4-back, or maximum average COG-ED *d'* of 4,48, given all perfect hits, and no misses or false alarms. Equally a theoretical minimum average COG-ED *d'* of -4,48, given no hits and all false alarms.

For the second phase the main measurement of interest was a participants effort-reward threshold or indifference point (*COG-ED IP)*. This was calculated by averaging their indifference points in the three experimental runs. For each of the three blocks (1-back v. 2-, 3-, and 4-back) the theoretical 7th offering for the 1-back would be their indifference point. A participant always choosing the higher N-back, would be offered (for the 1-back): 1.00, then 1.50, then 1.75, then 1.88, then 1.94, then 1.96, and the seventh 'offering' of 1.99 would be their indifference point in that block. A participant always choosing the 1-back, would be offered (for the 1-back): 1.00, then 0.50, then 0.25, then 0.12, then 0.06, then 0.03, and the seventh 'offering' of 0.01 would be their IP in that block. Yielding a theoretical maximum average COG-ED IP of 1.99, all high N-back chosen in all blocks on all levels, and a theoretical minimum average COG-ED IP of 0.01, all 1-back chosen in all blocks on all levels.

Exclusion. Participants were excluded task-wise according to the pre-registered exclusion criteria: lower than 80% accuracy on the DST-task, indicated knowledge of more than half of the RQ-items (RQ-Db response of 4 or 5) or familiarity with the COG-ED task. Participants with missing data were excluded task-wise. All other responses were kept in, including from drop-outs.

Exclusions were done prior to all descriptives and analyses.

Summary statistics. Summary descriptives were calculated for all tasks. A large portion of our tasks and measurements were ordinal in nature, and some of them had normality and/or homogeneity of variance violations, as well as outliers. As we are interested in *participants'* scoring differently on the different tasks, and not the relationship between the task-responses and scales themselves (e.g., Field, 2012), no outliers were excluded, no transformation attempts were made, and non-parametric tests were the norm, due these reasons medians and quartiles will be presented for the summary descriptives.

Internal reliability for the tasks and measurements were calculated using Cronbach's alpha. For the DST-tasks' measurement of willingness to exert cognitive effort: by using each block's low demand preference (LDP) as an item. Kool et al. (2010) found this to be high (Cronbach's alpha = .85). For the NfC-measurements: across all items (after reversing the 'low need for cognition'-items). For the COG-ED's measurement of willingness to exert cognitive effort: by using each phase two block's IP as an item, and for the COG-ED's measurement of intellectual ability (IQ/working-memory approximation): by using each N-back level's d' as an item. For the RQ-task: across all items. For the exploratory RQ-variables: across the non-heuristic items (RQ-nH) and across the heuristic items (RQ-H), and across the CRT-items (RQ-H, excluding the probability matching item[item 3]).

Debrief analyses. The DST Debrief answers (DST-d) were subjectively interpreted, had non-exclusive categories, and post-hoc rationalization and demand characteristics from the participants were a possibility (simply stating "yes" on the form, even though they didn't actually catch the manipulation), thus any analyses based on these variables must be treated with high caution. Kool et al. (2010) found that an awareness of the manipulation didn't influence the low demand preference across participants. To explore this we ran Mann-Whitney U test's on the low demand preference (DST1-LDP and DST2-LDP) between participants that caught the manipulation in the specific session (DST1-d and DST2-d scores of 1, '0.5' scores were recoded '0' for these tests), with Spearman's rank correlation as a post-hoc test on significant results. Other than this only frequencies of these results were calculated, as the main objectives with this debrief was to explore how many of the participants caught the manipulation in session two compared to (a cautious estimate) in session one, as well as the frequency of technical based preferences reported.

The RQ Debrief question were mainly used to detect any to be excluded (due to knowledge of most or all items), as research indicates that although familiarity with the CRT-items are common and raises the raw score, it doesn't affect the predictive value of the CRT-items (Mialek & Pennycook, 2018). To explore this we ran Mann-Whitney U test's on our seven main variables, and four exploratory variables, based on no prior experience (RQ-d = 0) or any prior experience (RQ-d > 0). However as the RQ Debrief question didn't target the CRT-items only, and the sub-group *n* were small, these results must be treated with caution.

Confirmatory hypotheses testing.

Hypothesis 1: Willingness to exert cognitive effort and deliberate reasoning. We predicted the correlation between DST1-LDP and RQ to be -.50, a higher avoidance of cognitive effort (i.e., high 'low demand preference') would be associated with a lower score on the RQ-task. The correlation between DST1-LDP and RQ was calculated in a Spearman's rank correlation coefficient analysis.

Hypothesis 2: Willingness to exert cognitive effort and thinking disposition. We predicted the correlation between DST1-LDP and NfC1 to be -.50, a higher avoidance of cognitive effort (i.e., high 'low demand preference') would be associated with a lower measure of 'Need for Cognition'. The correlation between DST1-LDP and NfC1 was calculated in a Spearman's rank correlation coefficient analysis.

Hypothesis 3: Willingness to exert cognitive effort - DST and COG-ED. We predicted the correlation between DST2-DLP and COG-ED IP to be -.50, a higher avoidance of cognitive effort (i.e., high 'low demand preference') would be associated with a lower indifference point (i.e, threshold between effort and reward). The correlation between DST2-LDP and COG-ED IP was calculated in a Spearman's rank correlation coefficient analysis.

Hypothesis 4: Test-retest reliability of the DST-task. We predicted the test-retest correlation between DST1-LDP and DST2-LDP to be .50, a higher low demand preference in one session would be associated with a higher low demand preference in the other session. The correlation between DST1-LDP and DST2-LDP was calculated in a Spearman's rank correlation coefficient analysis.

Secondary hypotheses testing. A correlation table with all main variables were computed using Spearman's rank correlation coefficients. Of particular interest were:

- RQ x COG-ED IP: To explore if participants' willingness to exert effort for reward correlated with their deliberate reasoning.
- NfC2 x COG-ED IP: To explore if participants' 'Need for Cognition' correlated with their effort/reward threshold.
- COG-ED *d*' x COG-ED IP: To explore if participants' executive ability correlated with their effort/reward threshold, as Westbrook et al. (2013) found this correlation to be . 32.
- NfC1 x NfC2: The test-retest reliability of NfC is usually found to be high (e.g., test-retest correlation of .88 with 7 weeks in between; Sadowski & Gulgoz, 1992).
- COG-ED *d*' x DST-LDP2: To explore if participants' executive ability correlated with their low demand preference. Kool et al. (2010) found a negative correlation of -.54 between switch cost (as an approximation of executive control) and low demand preference.

To examine the relationship between willingness to exert cognitive demand, thinking disposition and deliberate reasoning, a mediation analysis was carried out, with DST1-LDP as the predictor and RQ as the dependent, using NfC1 as a mediator.

To examine the relationship between intellectual ability, thinking disposition and deliberate reasoning, a rank-order moderation analysis was carried out, with NfC (average of NfC1 and NfC2 as the other tasks are from both sessions) as the predictor, RQ as the dependent, and COG-ED *d'* as the moderator. The effectiveness of our prediction model was examined by computing Spearman's rank correlation coefficient between our predicted RQ-values and the observed RQ-values.

Post-hoc exploratory testing. The four exploratory variables (RQ-H, RQ-nH, RQ-HRS and CRT) were included in the correlation table, computed using Spearman's rank correlation coefficients.

To examine the relationship between willingness to exert cognitive demand, thinking disposition and deliberate reasoning in both heuristic and non-heuristic items, mediation analyses was carried out, with DST1-LDP as the predictor and one with RQ-H as the dependent and one with RQ-nH as the dependent, using NfC1 as a mediator.

To examine any differences in thinking disposition and executive ability interaction between the heuristic and non-heuristic items, two moderation analyses were carried out. One using the heuristic items, RQ-H as the dependent variable, and one using the non-heuristic items, RQ-nH as the dependent variable, with NfC (average of NfC1 and NfC2 as the other tasks are from both sessions) as the predictor and COG-ED *d'* as the moderator in both instances, using ranked order moderation analyses. The effectiveness of our prediction models was examined by computing Spearman's rank correlation coefficient between our predicted RQ-H values and the observed RQ-H values, and between our predicted RQ-nH values and the observed RQ-nH values.

Results

Exclusion and Drop-out

The following data were excluded: In the DST1-task one participant was excluded due to low accuracy (DST1-ACC = .49), this participant was also excluded from the NfC1-measurement due to missing data, and dropped out of the project between session one and session two. In the DST2-task one participant was excluded due to low accuracy (DST2-ACC = .49). In the RQ-task one participant was excluded after indicating familiarity with more than half of the RQ-items (RQ-Db = 5).

No participants were excluded or prevented from participating due to our 'psychology-student'-limit of 50%. Six participants reported being psychology-students, 15% of our total sample.

Summary Statistics

Summary descriptive statistics for the seven main variables, and the four exploratory variables are presented in Table 1.

Table 1

Summary statistics with location, dispersion, shape of the distribution and internal consistency for the measurements of willingness to exert cognitive effort, thinking disposition, intellectual ability and deliberate reasoning

					Range			
Variable	*n*	*Mdn*	IQR (25% - 75%)	α	Potential	Actual	Skew (SE)	Kurtosis (SE)
DST1-LDP	39	.51	.10 (.48 - .58)	.71	0 - 1	.42 - .76	1.34 (0.38)	1.65 (0.74)
DST2-LDP	38	.51	.11 (.49 - .60)	.52	0 - 1	.31 - .75	0.39 (0.38)	0.47 (0.75)
NfC1	39	62	18 (53 - 71)	.83	18 - 90	38 - 81	-0.26 (0.38)	-0.62 (0.74)
NfC2	39	65	18 (55 - 73)	.89	18 - 90	34 - 81	-0.74 (0.38)	0.02 (0.74)
COG-ED IP	39	1.16	0.99 (0.83 - 1.82)	.79	0.01 - 1.99	0.48 - 1.98	0.12 (0.38)	-1.50 (0.74)
COG-ED *d*'	39	2.37	0.60 (2.11 - 2.71)	.67	-4.48 - 4.48	1.15 - 3.76	0.30 (0.38)	0.59 (0.74)
RQ	39	7	3 (5 - 8)	.65	0 - 14	1 -11	-0.38 (0.38)	-0.37 (0.74)
RQ-nH	39	3	2 (2 - 4)	.26	0 - 7	0 - 6	0.11 (0.38)	0.14 (0.74)
RQ-H	39	4	2 (3 - 5)	.63	0 - 7	0 - 7	-0.31 (0.38)	-0.29 (0.74)
RQ-HRS	39	5	2 (4 - 6)	-	0 - 7	1 -07	-0.46 (0.38)	0.01 (0.74)
CRT	39	4	2 (2 - 4)	.67	0 - 6	0 - 6	-0.38 (0.38)	-0.50 (0.74)

Note. Sample size (n), Median (Mdn), Interquartile range (IQR) with lower and upper quartile, Cronbach's alpha (α), Potential and Actual Range, Skewness (Skew) with standard error, and Kurtosis with standard error for the seven main variables, and the four exploratory variables. Willingness to exert cognitive effort as measured by DST's low demand preference (DST*-LDP) and COG-ED's Indifference Points (COG-ED IP). Thinking disposition as measured in Need for Cognition (NfC*), Intellectual ability as executive ability / working memory as measured by COG-ED's N-back d' (COG-ED *d*'). Deliberate reasoning as measured by RQ task, sub-divided into non-heuristic (RQ-nH), heuristic (RQ-H) and Cognitive Reflection Test (CRT) subcategories. Heuristic response suppression as measured by all non-heuristic responses in RQ-H (RQ-HRS).

Debrief

DST Debriefs. Frequencies can be seen in Table 2. Median low demand preference for those who explicitly reported finding the manipulation and those who didn't were .53 and .51 in session one, and .58 and .52 in session two, and the distributions differences were tested in session one (Mann-Whitney $U = 50.0$, $n_1 = 5$, $n_2 = 34$, $p = .146$) and in session two (Mann-Whitney $U = 78.5$, $n_1 = 13$, $n_2 = 25$, $p =.010$). A follow-up Spearman's rank correlation was run to assess the relationship between explicitly stated manipulation discovery and low demand preference in session two, and there was a medium-strong positive correlation between discovery and increased low demand preference, $r_s(36) = .43$, $p = .008$.

Table 2

Frequencies of manipulation discovery and other reported cue preferences in the DST-tasks.

	Frequency		Percent	
Variable[respons]	DST1 (*n* = 39)	DST2 (*n* = 38)	DST1	DST2
DST-Db[1]	5	13	12.82	34.21
DST-Db[0.5]	6	9	15.38	23.68
DST-Db[0]	28	16	71.79	42.11
DST-Tp[1]	10	13	25.64	34.21
DST-Up[1]	10	7	25.64	18.42

Note. Manipulation discovery (DST-Db; 0.5, partial; 1, full), technical preferences (DST-Tp) and unrelated preferences (DST-Up) as reported in the DST Debriefs. The three variables are non-exclusive.

RQ Debrief. 25 participants reported they had no prior experience with any of the RQ-items, 11 participants reported prior experience with a few of the RQ-items (less than four items), and three participants reported prior experience with almost half of the items (in addition to the aforementioned excluded participant with knowledge of most or all of them). Median RQ-score for those who reported prior experience and those who reported no prior experience were 6.0 and 8.5, and the distribution difference were tested in a Mann-Whitney $U = 44.0$, $n_1 = 25$, $n_2 = 14$, $p < .001$. A follow-up Spearman's rank correlation was run to assess the relationship between any prior experience with some of the RQ-items and the RQ-score. There was a strong positive correlation between prior experience and RQ-score, $r_s(36) = .63$, $p < .001$. For the other variables, no meaningful differences were found, except for the NfC1 and NfC2. Median NfC1-, and NfC2-scores for those who reported prior experience and those who reported no prior experience with any of the RQ-items were 70.5 and 56.5 in session 1, and 72.0 and 60.5 in session two, and the distribution differences were tested in session one (Mann-Whitney $U = 67.5$, $n_1 = 24$, $n_2 = 14$, p = .002) and in session two (Mann-Whitney $U = 63.5$, $n_1 = 24$, $n_2 = 14$, p = .002). Follow-up Spearman's rank correlation was run to asses the differences in session one, $r_s(36) = .50$, $p = .001$, and in session two, $r_s(36) = .52$, $p < .001$.

Analyses

Confirmatory hypotheses testing. Hypothesis 1. A Spearman's rank-order correlation between deliberate reasoning score as measured in the RQ-task, and avoidance of exerting cognitive effort as measured in the DST (DST1-LDP) resulted in a small-medium negative correlation, $r_s(37) = -.37$, $p = .011$ (one-tailed).

Hypothesis 2. A Spearman's rank-order correlation between avoidance of cognitive effort as measured in the DST (DST1-LDP), and thinking disposition as measured in the 'Need for Cognition' (NfC1) resulted in a medium negative correlation, $r_s(38) = -.50$, $p < .001$ (one-tailed).

Hypothesis 3. A Spearman's rank-order correlation between avoidance of cognitive effort, as measured in DST (DST2-LDP) and effort/reward threshold as measured in COG-ED (COG-ED IP) resulted in a small negative correlation, $r_s(37) = -.24$, $p = .069$ (one-tailed).

Hypothesis 4. A Spearman's rank-order correlation between avoidance of cognitive effort as measured in DST in session one (DST1-LDP), and avoidance of cognitive effort as measured in DST in session two (DST2-LDP) resulted in a medium positive correlation, $r_s(37) = .54, p < .001$ (one-tailed).

Secondary hypotheses testing. All intercorrelations between the seven main variables and four exploratory variables can be seen in Table 3, using Spearman's rank-order correlation coefficients, of the ones with particular interest denoted.

The relationship between avoidance of cognitive effort (DST1-LDP) and deliberate reasoning (RQ) was mediated by thinking disposition (NfC1). The standardized regression coefficient between DST1-LDP and NfC1 was -0.42, $p = .008$, and the standardized regression coefficient between DST1-LDP and RQ was -0.25, $p = .133$, and the standardized indirect effect between NfC1 and RQ, when controlling for DST1-LDP was 0.37, $p = .039$. Lowering the standardized regression coefficient between DST1-LDP and RQ to -0.09, $p = .601$. We tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 5,000 samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th percentile and 97.5th percentile. The bootstrapped unstandarized indirect effect was 0.096, and the 95% confidence interval ranged from 0.022 to 0.181.

Intellectual ability as measured with COG-ED *d'* was examined as a moderator of the relationship between thinking disposition (average of NfC1 and NfC2) and deliberate reasoning (RQ). In a rank-order regression analysis with RQ as the dependent, and NfC and COG-ED *d'* as predictors, the unstandardized regression coefficients for NfC was 0.093, $p = .011$ and for COG-ED *d'* was 0.518, $p = .483$. In a rank-order regression analysis with RQ as the dependent, and NfC, COG-ED *d'* and interaction(COG-ED *d'* and NfC) as predictors, the unstandardized regression coefficients for NfC was -0.049, $p =.693$, and for COG-ED *d'* was -4.07, $p = .284$, and for the interaction was 0.072, $p = .205$. A Spearman's rank-order correlation between our predicted RQ values and the observed RQ values resulted in a medium positive correlation, $r_s(36) = .59$, $p < .001$.

Table 3

Summary of intercorrelations between willingness to exert cognitive effort, thinking disposition, intellectual ability and deliberate reasoning.

Variable	Statistic	1	2	3	4	5	6	7	8	9	10
1. DST1-LDP	r_s										
	p value										
2. DST2-LDP	r_s	.54[a]									
	p value	.005									
3. NfC1	r_s	-.50[a]	-.39								
	p value	.001	.017								
4. NfC2	r_s	-.43	-.21	.82[b]							
	p value	.007	.199	<.001							
5. COG-ED IP	r_s	-.25	-.25[a]	.04	.09[b]						
	p value	.122	.138	.806	.600						
6. COG-ED *d*'	r_s	-.03	.03[b]	.18	.25	.25[b]					
	p value	.853	.842	.269	.129	.122					
7. RQ	r_s	-.37[a]	-.11	.46	.57	.01[b]	.32				
	p value	.023	.507	.003	<.001	.954	.047				
8. RQ-nH[c]	r_s	-.42	-.07	.52	.52	-.12	.17	.80			
	p value	.009	.699	.001	.001	.462	.311	<.001			
9. RQ-H[c]	r_s	-.28	-.18	.31	.43	.13	.37	.88	.45		
	p value	.084	.298	.055	.007	.424	.022	<.001	.004		
10. RQ-HRS[c]	r_s	-.43	-.12	.28	.37	.07	.24	.76	.52	.76	
	p value	.007	.490	.093	.022	.686	.140	<.001	.001	<.001	
11. CRT[c]	r_s	-.26	-.20	.27	.39	.14	.34	.82	.38	.96	.71
	p value	.121	.227	.101	.016	.412	.038	<.001	.016	<.001	<.001

Note. Willingness to exert cognitive effort as measured by DST's low demand preference (DST*-LDP) and COG-ED's Indifference Points (COG-ED IP).Thinking disposition as measured in Need for Cognition (NfC*). Intellectual ability as executive ability / working memory as measured by COG-ED's N-back d' (COG-ED *d*'). Deliberate reasoning as measured by RQ task, sub-divided into non-heuristic (RQ-nH), heuristic (RQ-H) and Cognitive Reflection Test (CRT) subcategories. Heuristic response suppression as measured by all non-heuristic responses in RQ-H (RQ-HRS). Intercorrelations for variables 7.-11. must be treated with caution as several are calculated on each other (e.g., 7. RQ is the sum of 8. RQ-nH and 9. RQ-H; 11. CRT is a sub-category of 9. RQ-H).

[a] Confirmatory analyses

[b] Secondary analyses of particular interest

[c] Post-hoc exploratory variable

Post-hoc exploratory testing. All intercorrelations between the seven main variables and four exploratory variables can be seen in Table 3, using Spearman's rank-order correlation coefficients.

The relationship between avoidance of cognitive effort (DST1-LDP) and deliberate reasoning in both heuristic-, and non-heuristic items was mediated by thinking disposition (NfC1).

The standardized regression coefficient between DST1-LDP and NfC1 was -0.42, *p* = .008, and the standardized regression coefficient between DST1-LDP and RQ-H was -0.16, *p* = .330, and the standardized indirect effect between NfC1 and RQ-H, when controlling for DST1-LDP was 0.26, *p* = .160. Lowering the standardized regression coefficient between DST1-LDP and RQ-H to -0.05, *p* = .778. We tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 5,000 samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th percentile and 97.5th percentile. The bootstrapped unstandarized indirect effect was 0.046, and the 95% confidence interval ranged from -0.008 to 0.109.

The standardized regression coefficient between DST1-LDP and NfC1 was -0.42, *p* = .008, and the standardized regression coefficient between DST1-LDP and RQ-nH was -0.28, *p* = .083, and the standardized indirect effect between NfC1 and RQ-nH, when controlling for DST1-LDP was 0.39, *p* = .024. Lowering the standardized regression coefficient between DST1-LDP and RQ-nH to -0.11, *p* = 498 We tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 5,000 samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th percentile and 97.5th percentile. The bootstrapped unstandarized indirect effect was 0.050, and the 95% confidence interval ranged from 0.010 to 0.093.

Intellectual ability as measured with COG-ED *d'* was examined as a moderator of the relationship between thinking disposition (average of NfC1 and NfC2) and deliberate reasoning in heuristic items (RQ-H) and non-heuristic items (RQ-nH).

In a rank-order regression analysis with RQ-H as the dependent, and NfC and COG-ED *d'* as predictors, the unstandardized regression coefficients for NfC was 0.045, *p* = .057 and for COG-ED *d'* was 0.762, *p* = .121. In a rank-order regression analysis with RQ-H as the dependent, and NfC, COG-ED *d'* and interaction(COG-ED *d'* x NfC) as predictors, the unstandardized regression coefficients for NfC was 0.050, *p* =.518, and for COG-ED *d'* was 0.917, *p* = .696, and for the interaction was -0.003, *p* = .941.

In a rank-order regression analysis with RQ-nH as the dependent, and NfC and COG-ED *d'* as predictors, the unstandardized regression coefficients for NfC was 0.041, *p* = .035 and for COG-ED *d'* was 0.152, *p* = .703. In a rank-order regression analysis with RQ-nH as the dependent, and NfC, COG-ED *d'* and interaction(COG-ED *d'* and NfC) as predictors, the unstandardized regression coefficients for NfC was -0.053, *p* =.308, and for COG-ED *d'* was 2.658, *p* = .094, and for the interaction was 0.044, *p* = .062. A Spearman's rank-order correlation between our predicted RQ-H values and the observed RQ-H values resulted in a medium positive correlation, $r_s(36) = .38$, *p* = .020. A Spearman's rank-order correlation between our predicted RQ-nH values and the observed RQ-nH values resulted in a medium positive correlation, $r_s(36) = .57$, *p* = .002.

Discussion

We hypothesized that willingness to exert cognitive effort as measured in the DST would show moderate reliability in a retest four to eight weeks later, go well together with the same concept as measured in COG-ED, and that it should be well captured by the self-reported NfC. Both the reliability (.54), and correlation with NfC (-.50), was as hypothesized, however the relationship with the measurement from the COG-ED task was weaker than predicted (-.25). In addition, subsequent exploratory analyses on the DST results casts some concerns on the construct validity of this measurement. We will take a look at each of these to examine what these results imply for the DST measurement, and for our use of this measurement in relation to deliberate reasoning.

Our main hypothesis was that a willingness to exert effort would increase successful deliberate reasoning on tasks both with and without an intuitive-, gut-answer. While we did find an overall effect of willingness to exert effort on deliberate reasoning, this effect was more modest than predicted and only explained 14% of the variance of participants' ranking on deliberate reasoning. Further analyses using NfC as a measurement of willingness to exert effort (through an increased disposition towards cognitively effortful thinking) and COG-ED *d*' as a measurement of intellectual ability / working memory approximation, explained 35% of the variance of participants' ranking on deliberate reasoning. However exploratory analyses reviled that this might mostly be due to increased scores on the non-heuristic items in the deliberate reasoning task, and less so on the heuristic items.

These findings must be explored to answer our overall questions of *why* and *when* we rely on intuitive responses instead of engaging in deliberate reasoning.

The 'Why' - Willingness to exert Cognitive Effort

As system 2's deliberate reasoning is often assumed to come at a cost, a cost we find aversive (e.g., Kool et al., 2010), we needed a measurement of willingness to exert cognitive effort as this should be predictive of when an individual engage in cognitively demanding processing. To explore this we used two different experimental paradigms, the Demand

Selection Task (Kool et al., 2010) and the Cognitive Effort Discounting Paradigm (Westbrook et al., 2013), as well as measured participants' thinking disposition, through their self-reported tendency to enjoy, or engage in complex thinking using the Need for Cognition scale (Cacioppo, Petty, & Kao, 1984; short version). We hypothesized that these should capture the same concept, and that the DST's measurement of willingness to exert cognitive effort would show moderate reliability in a retest four to eight weeks later.

The test-retest was as hypothesized, with a correlation of .54. However whereas Kool et al. (2010) found an average low effort demand preference of .67 (.50 would indicate random/chance cue preference; in Experiment 5) our results did not mirror this, with .53 in session one and .54 in session two. Kool et al. (2010) also found that an awareness of the manipulation did not influence the low effort demand preference, and while we did not administer a full debrief after session one, in session two 34% of our participants reported finding the manipulation, and these showed higher effort avoidance than the unaware participants. In addition our internal consistency in assessing preference for the low effort demand cue across the task was far lower than in Kool et al. (2010), where they found this to be .91, we found it to be .71 in session one and dropping down to .52 in session two. Even though these results indicate that the DST does not retest well, it does not necessarily imply that the first assessment didn't measure a willingness to exert cognitive effort. Each DST assessment consisted of 600 trials using the same two easy tasks in each trial, making it rather tedious, and this could have affected the retest.

The predicted correlation between the DST's low effort demand preference and COG-ED's reward/effort-threshold was -.50, however a -.25 correlation was observed (for both sessions of the DST). There are multiple plausible explanations for this, two immediate ones are the differences in intrinsic versus extrinsic reward and implicit versus explicit effort demands.

The DST provides participants with no indication of their performance, nor any rewards of any kind, virtual points or otherwise. The COG-ED however provides participants with both a virtual intermediate extrinsic reward (their performance and running-total), and an actual extrinsic reward after task completion. And whereas the DST relies on an implicit manipulation of effort demand, the COG-ED clearly states the effort demands and the participant are then to choose between a low or high effort task. Kool et al. (2010) found in another implicit task-paradigm that an extrinsic reward significantly dampened the avoidance of cognitive effort. This in line with research by Sandra and Otto (2018) and others (Thompson, Chaiken, & Hazlewood, 1993) who demonstrated that the use of external reward undermine the intrinsic willingness to exert effort, particularly in participants reporting a lower 'Need for Cognition'. Something our research indicates as well with no relation between COG-ED's measurement of willingness to exert effort and participants self-reported thinking disposition (NfC).

Another major difference between the two paradigms is in the task demands themselves. The DST uses two very easy tasks, and the effort demand manipulation is in the slight effort associated with switching between these two tasks. In the COG-ED however, the task itself is very demanding especially at the higher N-back levels. Both this present study and in Westbrook et al (2013) it was found that increased executive ability increased the participant's effort-reward threshold. Interestingly while Kool et al. (2010) found that participants' low effort demand was negatively correlated (-.54) with their executive function approximation (as measured by switch cost in a 'no cue choice' block with predetermined task switching sequence), no such relation was found in the present study using the executive ability as measured in the N-back phase of COG-ED.

Taken together these results makes this weak correlation between the two measures of willingness to exert effort less surprising. The COG-ED's use of reward and heavily demanding task might have undermined a participant's intrinsic willingness to exert effort, and instead reflect differences in cognitive ability and reward-sensitivity. And although participants did 'compete' against their own performance in the different N-back levels, the subjectively experienced demand in the different levels might have been mainly driven by cognitive ability and less so the fact that they only had to perform at their own 'level'. This does not necessarily imply that the COG-ED doesn't measure an overall aversiveness to effort, however it lowers the usefulness of it for our purpose of individual predictions on intrinsically motivated tasks.

This leave us with one important question, do the DST measure willingness to exert cognitive effort? In accordance with our hypothesis, the DST's measure of avoidance of cognitive effort did correlate well (-.50) with thinking disposition (NfC) however this does not necessarily imply that the DST actually measures a participant's willingness to exert effort. As mentioned we found no relationship between executive ability as measured in the N-back phase of COG-ED and the DST's measure of willingness to exert cognitive demand. This could be due to the fact that our measure of executive ability is more of a maximum capacity of the algorithmic mind, and that the DST rather relies on the experienced effort demands imposed by a cognitive flexibility requirement (see Haatveit et al., 2010). However, this finding is noteworthy and should be addressed in future research by comparing the DST to other measures of executive function and measures of individual experienced effort demand. In addition our overall demand preference did not mirror that of Kool et al. (2010)

in that it did not show a meaningful difference from chance, and upon manipulation detection participants seemed to favor the low demand preference in session two. However as the DST might not retest well, and we used a conservative partial debrief for session one, these results must be interpreted with caution. It does however posit the possibility that the DST's correlation with NfC is due to some other factor(s), and not wholly attributable to an aversiveness of the effort required in the task. Perhaps participants high in 'Need for Cognition' might be more bored, and thus switch between the cues more often while participants low in 'Need for Cognition' might stay at one cue for longer periods, leading to differences in if-, or how fast, they found the manipulation. As we only included a partial debrief in session one, and our results indicate that the DST might not retest well, we couldn't explore this cue-switching hypothesis further, but it warrants some additional analyses and caution going forward.

Taken together these results indicate that while the notion of effort aversion do carry high face validity, and have been explored more thoroughly in the two original experimental paradigms, our findings cast some doubt on the certainty we can claim this to be demonstrated in this present study, especially on an individual level.

The 'When' - Individual differences in Critical Thinking

Our main hypothesis was that a willingness to exert effort would increase successful deliberate reasoning on tasks both with and without an intuitive-, gut-answer. While we did find an overall effect of willingness to exert effort on deliberate reasoning, this effect was more modest than predicted and only explained 14% of the variance of participants' ranking on deliberate reasoning.

Subsequent analyses on the willingness to exert cognitive effort measurement prompted us to run a mediation analysis, and the correlation between our measure of willingness to exert effort and deliberate reasoning was mediated through thinking disposition. As we can't be sure if this is due to an actual mediation, i.e. all willingness to exert effort captured was due to thinking disposition, or if our measurement of willingness to exert effort didn't actually measure this willingness to exert effort *through effort*, we chose to use the 'Need for Cognition' measurement in our further exploratory analyses.

In our full model, using thinking disposition and executive ability, we could predict 35% of the rank-order variance in the full deliberate reasoning task, however this explained variance was only 14% for the heuristic items. In addition we saw an interaction effect between thinking disposition and executive ability in the non-heuristic items, in which participants' positive effect of thinking disposition was positively influenced by higher executive ability. While no such relationship could be seen in the heuristic items, where both high 'Need for Cognition' and high executive ability contributed positively towards successful deliberate reasoning in heuristic items, but with no interaction between them. This can reflect differences in the difficulty in the non-heuristic and heuristic items (given detection), or an actual difference in the effects of these aspects on detection, or be an artifact due to a higher NfC in the participants that indicated prior experience with our RQ-items (and scored higher). However, as our RQ-task contained a mix of both heuristic and non-heuristic items, this might have decreased the advantage of prior knowledge, as a participant wouldn't know it's a 'trick questions' section throughout, but rather need to find the conflict in all items they did not have prior experience with.

Taken together, these findings indicate that individuals' thinking disposition and executive ability do contribute somewhat to deliberate reasoning in tasks with a strong intuitive-, gut-answer, however these effects are modest. In deliberate reasoning tasks without such an intuitive response however these contributions are far greater, especially for people high in both executive ability and with a high tendency towards enjoying, or engaging in complex thinking.

Implications and Future Directions

This research highlights the need for further research in particularly four different areas.

A more robust measurement of willingness to exert cognitive effort in individuals. While the present study attempted to measure this willingness to exert cognitive effort, this was partly unsuccessful. A reliable measurement of willingness to exert cognitive effort is needed, either through a verification of the DST paradigm by further examination of it and linking it to other measures of intellectual ability (to show that it does indeed measure cognitive effort avoidance) or through development of new measurements. As effort aversion continues to be an often used explanation of why we rely on intuitive reasoning, a measurement for this is needed in order to link these two concepts.

A broader range of heuristic/conflict detection tasks. While prior experience might not lower the predictive value of the CRT-items, it does complicate the interpretation when attempting to examine the underlying mechanisms behind successful deliberate reasoning in tasks with a heuristic response. Dividing participants into a prior experience group and a non-prior experience group might not suffice as individual factors might influence who have sought out (or remembers) the items.

Further research into the role of conflict detection and suppression in deliberate reasoning. While able to predict some of the variance in participants' ranking on the deliberate reasoning task, this was less so for the heuristic items. Indicating that a large portion of the variance is due to conflict detection between the heuristic response and the correct response. Supported by our results that show that participants either gave the correct or the heuristic answer, and much less frequent any other (incorrect) answer. The exact process of this whether it's purely a conflict detection or both a conflict detection and a sustained 'decoupling' that is needed is still debated (e.g., De Neys, Vartanian, & Goel, 2008; De Neys, Rossi, & Houdé, 2013; De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2015b), though there has been some evidence that the two systems operate in, or are especially dependent on, somewhat different parts of the brain (e.g., Tsujii & Watanabe, 2009; Aron, Robbins, & Poldrack, 2014; De Nyes et al, 2008). A thorough understanding of this process is needed as it seems to be highly important in predicting when successful deliberate reasoning takes place in tasks requiring a cognitive reflection ability.

The 'state or trait'-ness of critical thinking. In the present study, a wide range of 'common' cognitive aspects was measured, both executive ability, thinking disposition and different measures of willingness to exert cognitive effort. The fact that neither of these seems to capture conflict detection and subsequent successful deliberate reasoning particularly well might indicate that this is influenced by other aspects than stable cognitive traits and abilities.

Limitations

There are multiple possible limitations in our findings, the most important ones are related to our sample.

We tested mainly students and while some research indicate that 'WEIRD'-samples generalize well (e.g., Hanel & Vione, 2016; western educated industrialized rich and democratic), our instruments might have been particularly sensitive as we measured thinking disposition, executive function, and performance on 'test-like' tasks. Unsuccessful deliberation due to capacity limitations in the algorithmic mind might have been less likely in our sample compered to the general population. In addition, intrinsic motivation to perform well might be higher in students, especially in experiments in which you know your answers will be scrutinized. This might also have lead to an increased chance of conflict detection for the heuristic responses, and an increased motivation to sustained override of these responses. Although this should be particularly true for participants high in 'Need for Cognition' and thus not pose a problem within the study itself.

Another limitation is our sample size. In the RQ-task debrief, 15 of our participants indicated prior experience with at least some of the items, and although we didn't differ between the heuristic and non-heuristic items in this debrief, experience with the heuristic items can be assumed to be the norm. Research indicates that while prior experience do increase the raw scores, it doesn't lower their predictive power (Bialek & Pennycook, 2018). However as we wanted to look at the proposed mechanisms behind these scores, i.e. predict them, it poses a possible limitation for this study, and while the sample size was large enough for the pre-registered hypotheses, it was to small to carry out the exploratory analyses on the two different sub-groups separately. Although we assume that a larger sample size wouldn't change our conclusions, as 60% of our naive participants got the first item correct, a heuristic item from the CRT, and it doesn't look like this group difference is due to a 'head-start' for participants with prior knowledge. Indicating that this knowledge-effect could be partially due to actual differences for the individuals who have either sought out these items previously, or remembers encountering them, and those who have not. As shown they did differ in NfC, and the NfC had a high test-retest reliability limiting the possibility that this measurement was influenced by the RQ-task.

Conclusion

In examining why and when we rely on intuitive responses in leu of more demanding deliberate reasoning, our findings indicate that one's disposition towards effortful thinking as well as cognitive ability do affect one's deliberate reasoning, in which 35% of the variance in our participants ranking could be explained. This however was especially true in tasks without an interfering and strong intuitive response. In the tasks with a strong intuitive response however, only 14% of the variance could be explained by one's thinking disposition and cognitive ability. Indicating that a detection and suppression of this intuitive response plays a very large role in successful deliberate reasoning. While most individuals might be able to carry out this deliberate reasoning, the critical factor is whether or not they detect the *need* for it. An implication supported by the fact that our participants on average only got half of the items correct, even though they were mostly students, making limitations due to cognitive ability less likely, and the items themselvs are not very demanding once you detect the conflict in the intuitive response.

This study also demonstrated some possible weakness in two different experimental paradigms measuring willingness to exert cognitive demand, especially on an individual level and for use in relation to intrinsic motivated deliberate reasoning.

These findings highlight the need for future research into detection and suppression of intuitive responses, as well as a need for a more robust measurement of willingness to exert cognitive demand, especially if intended for individual predictions.

References

American Management Association (2012). *AMA 2012 Critical skills survey*. Retrieved March 20, 2019, from https://playbook.amanet.org/wp-content/uploads/ 2013/03/2012-Critical-Skills-Survey-pdf.pdf

Ariely, D. (2010). *Predictably irrational: The hidden forces that shape our decisions.* New York: Harper Perennial.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: one decade on. *Trends in Cognitive Sciences, 18*, 177-185. doi:10.1016/j.tics. 2013.12.003

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods, 50*, 1953-1959. doi:10.3758/ s13428-017-0963-x

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433-436.

Bruinsma, J., & Crutzen, R. (2018). A longitudinal study on the stability of the need for cognition. *Personality and Individual Differences, 127*, 151-161. doi:10.1016/j.paid. 2018.02.001

Cacioppo, J., Petty, R. E., Feinstein, J., & Jarvis, W. B. J. (1996). *Dispositional differences in cognitive motivation: The life and times of individuals varying in Need for Cognition* (119).

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116-131. doi:10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of Need for Cognition. *Journal of Personality Assessment, 48*, 306-307. doi:10.1207/ s15327752jpa4803_13

Cacioppo, J. T., Petty, R. E., Kao, C. F., & Rodriguez, R. (1986). Central and peripheral routes to persuasion: An individual difference perspective. *Journal of Personality and Social Psychology, 51*, 1032-1043. doi:10.1037/0022-3514.51.5.1032

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*, 1248-1299. doi:10.1016/j.cognition.2007.06.002

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science, 19*, 483-489. doi:10.1111/j. 1467-9280.2008.02113.x

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review, 20*, 269-273. doi:10.3758/s13423-013-0384-5

EF EPI (2018). EF English Proficiency Index - Norway. In *EF.* Retrieved March 24, 2019, from: https://www.ef.no/epi/regions/europe/norway/

Evans, J. S. B. T. (1984). Heuristics and analytical processes in reasoning. *British Journal of Psychology, 75,* 451-468.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*, 223-241. doi: 10.1177/1745691612460685

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi:10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi:10.3758/BF03193146

Field, A. (2012). *Discovering statistics using R.* Los Angeles: SAGE

Fleischhauer, M., Enge, S., Brocke, B., Ullrich, J., Strobel, Al., & Strobel, An. (2009). Same or different? Clarifying the relationship of Need for Cognition to personality and intelligence. *Personality and Social Psychology Bulletin, 36*, 82-96. doi: 10.1177/0146167209351886

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25-42. doi:10.1257/089533005775196732

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*, 53-96. doi:10.1111/j.1539-6053.2008.00033.x

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75-90. doi:10.1037/0033-295X.109.1.75

Grahe, J. E. (2014). Announcing Open Science badges and reaching for the sky. *The Journal of Social Psychology, 154*, 1-3. doi:10.1080/00224545.2014.853582

Hanel, P. H., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PLOS ONE, 11*, e0168354. doi:10.1371/journal.pone.0168354

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock, P., & Meshkati, N. *Human Mental Workload* (p. 139-183). North Holland, Amsterdam.

Hull, C. L. (1943). *Principles of behavior: an introduction to behavior theory*. Oxford, England: Appleton-Century.

Haatveit, B. C., Sundet, K., Hugdahl, K., Ueland, T., Melle, I., & Andreassen, O. A. (2010). The validity of d prime as a working memory index: Results from the "Bergen n-back task". *Journal of Clinical and Experimental Neuropsychology, 32*, 871-880. doi: 10.1080/13803391003596421

Inquisit 5 Web (2018). [computer software]. Retrieved from hhttps://www.millisecond.com

JASP Team (2018). JASP (version 0.9.2). [computer software].

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Strauss and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-291. doi:10.2307/1914185

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin, 136*, 849-874. doi:10.1037/a0019842

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology, 55*, 352-358. doi:10.1037/h0043688

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36,* ECVP Abstract Supplement.

Kloke, J. D., & Mckean, J. W. (2012). Rfit: Rank-based estimation for linear models. *The R Journal, 4,* 57-64.

Koehler, J. D., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition, 38*, 667-676. doi:10.3758/MC.38.6.667

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*, 665-682. doi:10.1037/a0020198

Lau, R., & Redlawsk, D. (2001). Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science, 45,* 951-971. doi: 10.2307/2669334

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43*, 431-442. doi:10.1037/0003-066X.43.6.431

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence, 30*, 81-108. doi:10.1016/0004-3702(86)90068-8

Liefooghe, B. (2017). The contribution of task-choice response selection to the switch cost in voluntary task switching. *Acta Psychologica, 178*, 32-40. doi:https://doi.org/10.1016/j.actpsy.2017.05.006

MatLab (2018). MATLAB (version 9.4). [computer software]. The MathWorks, Natick, Massachusetts.

Melnikoff, D. E., & Bargh, J. A. (2018a). The insidious number two. *Trends in Cognitive Sciences, 22*, 668-669. doi:10.1016/j.tics.2018.05.005

Melnikoff, D. E., & Bargh, J. A. (2018b). The mythical number two. *Trends in Cognitive Sciences, 22*, 280-293. doi:10.1016/j.tics.2018.02.001

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E-J., Ware, J. J. & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. doi: 10.1038/s41562-016-0021

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600. doi:10.1073/pnas.1708274114

Nuzzo, R. (2014). Scientific method: Statistical errors. In *Nature*, Retrieved April 2, 2019, from https://www.nature.com/news/scientific-method-statistical-errors-1.14700

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General, 134*, 368-387. doi:10.1037/0096-3445.134.3.368

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25*, 46-59. doi:10.1002/hbm.20131

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making, 10*, 549-563.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition, 123*, 335-346. doi: 10.1016/j.cognition.2012.03.003

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science, 24*, 425-432. doi: 10.1177/0963721415604610

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015b). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34-72. doi:10.1016/j.cogpsych.2015.05.001

Pennycook, G., De Neys, W., Evans, J. S. B. T., Stanovich, K. E., & Thompson, V. A. (2018). The mythical dual-process typology. *Trends in Cognitive Sciences, 22*, 667-668. doi: 10.1016/j.tics.2018.04.008

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437-442.

Rips, L. J. (1989). The psychology of knights and knaves. *Cognition, 31*, 85-116. doi: 10.1016/0010-0277(89)90019-X

R Core Team (2018). R: A language and environment for statistical computing. [computer software]. R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team. (2016) RStudio: Integrated development for R. [computer software]. RStudio, Inc, Boston, MA.

Sadowski, C. J., & Gulgoz, S. (1992). Internal consistency and test-retest reliability of the Need for Cognition scale. *Perceptual and Motor Skills, 74*, 610-610. doi:10.2466/ PMS.74.2.610-610

Sandra, D. A., & Otto, A. R. (2018). Cognitive capacity limitations and Need for Cognition differentially predict reward-induced cognitive effort expenditure. *Cognition, 172*, 101-106. doi:10.1016/j.cognition.2017.12.004

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*, 99-118. doi:10.2307/1884852

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought.* New Haven, CT, US: Yale University Press.

Stanovich, K. E. (2016). The Comprehensive Assessment of Rational Thinking. *Educational Psychologist, 51*, 23-34. doi:10.1080/00461520.2015.1125787

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161-188. doi:10.1037/0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavior Brain Science, 23*, 645-665; discussion 665-726.

Stanovich, K. E., & West, R. F. (2014). The assessment of rational thinking: IQ ≠ RQ. *Teaching of Psychology, 41*, 265-271. doi:10.1177/0098628314537988

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking.* Cambridge: MIT Press

Teigen, K. H., & Keren, G. (2007). Waiting for the bus: When base-rates refuse to be neglected. *Cognition, 103*, 337-357. doi:10.1016/j.cognition.2006.03.007

Thaler, R. H. (2015). *Misbehaving: The making of behavioral economics.* New York: W W Norton & Co.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: improving decisions about health, wealth, and happiness.* New York: Penguin Books.

Thompson, E. P., Chaiken, S., & Hazlewood, J. D. (1993). Need for cognition and desire for control as moderators of extrinsic reward effects: A person × situation approach to the study of intrinsic motivation. *Journal of Personality and Social Psychology, 64*, 987-999. doi:10.1037/0022-3514.64.6.987

Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 94*, 197-209. doi:10.1037/0022-0663.94.1.197

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*, 1275. doi:10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*, 147-168. doi:10.1080/13546783.2013.844729

Treadway, M. T., Buckholtz, J. W., Schwartzman, A. N., Lambert, W. E., & Zald, D. H. (2009). Worth the 'EEfRT'? The effort expenditure for rewards task as an objective measure of motivation and anhedonia. *PLOS ONE, 4*, e6598. doi:10.1371/ journal.pone.0006598

Tsujii, T., & Watanabe, S. (2009). Neural correlates of dual-task effect on belief-bias syllogistic reasoning: A near-infrared spectroscopy study. *Brain Research, 1287*, 118-125. doi:10.1016/j.brainres.2009.06.080

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Vernier (n.d.). Hand Dynamometer (HD-BTA). [apparatus]. Vernier.

Wansink, B., & Sobal, J. (2007). Mindless eating: The 200 daily food decisions we overlook. *Environment and Behavior, 39*, 106-123. doi:10.1177/0013916506295573

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281. doi:10.1080/14640746808400161

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*, 129-133. doi: 10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “p < 0.05”. *The American Statistician, 73*(sup1), 1-19. doi: 10.1080/00031305.2019.1583913

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*, 930-941. doi:10.1037/a0012842

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLOS ONE, 8*, e68210. doi:10.1371/journal.pone.0068210

Wylie, G., & Allport, A. (2000). Task switching and the measurement of "switch costs". *Psychological Research, 63*, 212-233. doi:10.1007/s004269900003

Footnotes

[1] Roughly explained: the former is known as instrumental rationality, whether your actions are rational given an appropriate goal. The latter is known as epistemic rationality, and concerns the appropriateness of your goals.

[2] While 'intelligence', or more specifically fluid intelligence itself is within the algorithmic mind, our assessments of this intelligence relies on a motivation to work on these assessments, thus being influenced by our reflective mind (Stanovich, 2009).

[3] As rationality might refer to different types of rationality. Stanovich et al. (2016) makes the distinction that although named "The Rationality Quotient", this test mostly measures rational/critical thinking and concepts *needed* for rationality (in the broad sense). It does not measure actual rational choices made by the individual in their life.

[4] Due to either a typo, or misreading of the degrees of freedom as the suggested N, we went with a stopping criterion of N = 40, not the correct N = 41.

[5] No (potential) participant indicated not meeting the inclusion criteria. One potential participant inquired if a small concussion from several years ago would be an exclusion criteria, was told it was not, but chose not to participate. Another inquired if there were any prerequisite knowledge requirements, were told that it was not, and participated.

[6] The e-mail invited to a study within cognitive psychology, and included information of the estimated time-frame for each session, that there would be two test sessions, that they would be tested individually, the inclusion and exclusion criteria, the location of testing, the overall time-frame for the project, and the non-monetary compensation provided upon completion of both sessions. As these e-mail's were sent trough the study advisors, we did not have complete control over the exact content of the e-mail the participants received, however no participants received more information than described above.

[7] With the exception of the written instructions for the COG-ED task in session 2. Participants were encouraged to ask for an oral summary instruction if they pleased, however English fluency is very high in Norway (EF, 2019) and the instructions was aided with graphics.

[8] Based on casual feedback, this was understood by the participants. Most of the participants had either forgotten about the extra reward opportunity in session two, or expressed surprise that the extra-reward opportunity didn't concern a repeat of the EEfRT-task, but rather a new task (the COG-ED).

Appendix A

Informed consent

CREP, autumn 2018

INVITATION TO PARTICIPATE IN A RESEARCH PROJECT

DECISION-MAKING IN A CLASSIFICATION TASK

You are invited to participate in a research project that investigates how we classify stimuli and judge a range of statements.

WHO CAN PARTICIPATE?

People between the age of 18 and 50 with normal or corrected eyesight, without history of brain injury and/or surgery, and without any diagnosed neurological or mental health disorder in an acute stage can participate. Further, you should not have taken any central nervous system medication (e.g., antidepressants, antiepileptic drugs) or any recreational drugs (e.g. cannabis; excl. tobacco and alcohol) within the last 3 months.

The Department of Psychology at UiT The Arctic University of Norway is responsible for the implementation of this study.
Project Manager is Dr. Gerit Pfuhl, Associate Professor at the Department of Psychology, UiT, Tromsø (tel. [redacted], e-mail: [redacted]) and experiment leader is Kristoffer Klevjer, Masters student at the Department of Psychology, UiT, Tromsø (tel. 45054054, email: kkl012@uit.no) .

It is important that you read this information letter thoroughly before you agree to participate. Feel free to ask questions if you're wondering about anything by contacting Kristoffer Klevjer or Gerit Pfuhl.

WHAT IS THE STUDY ABOUT?

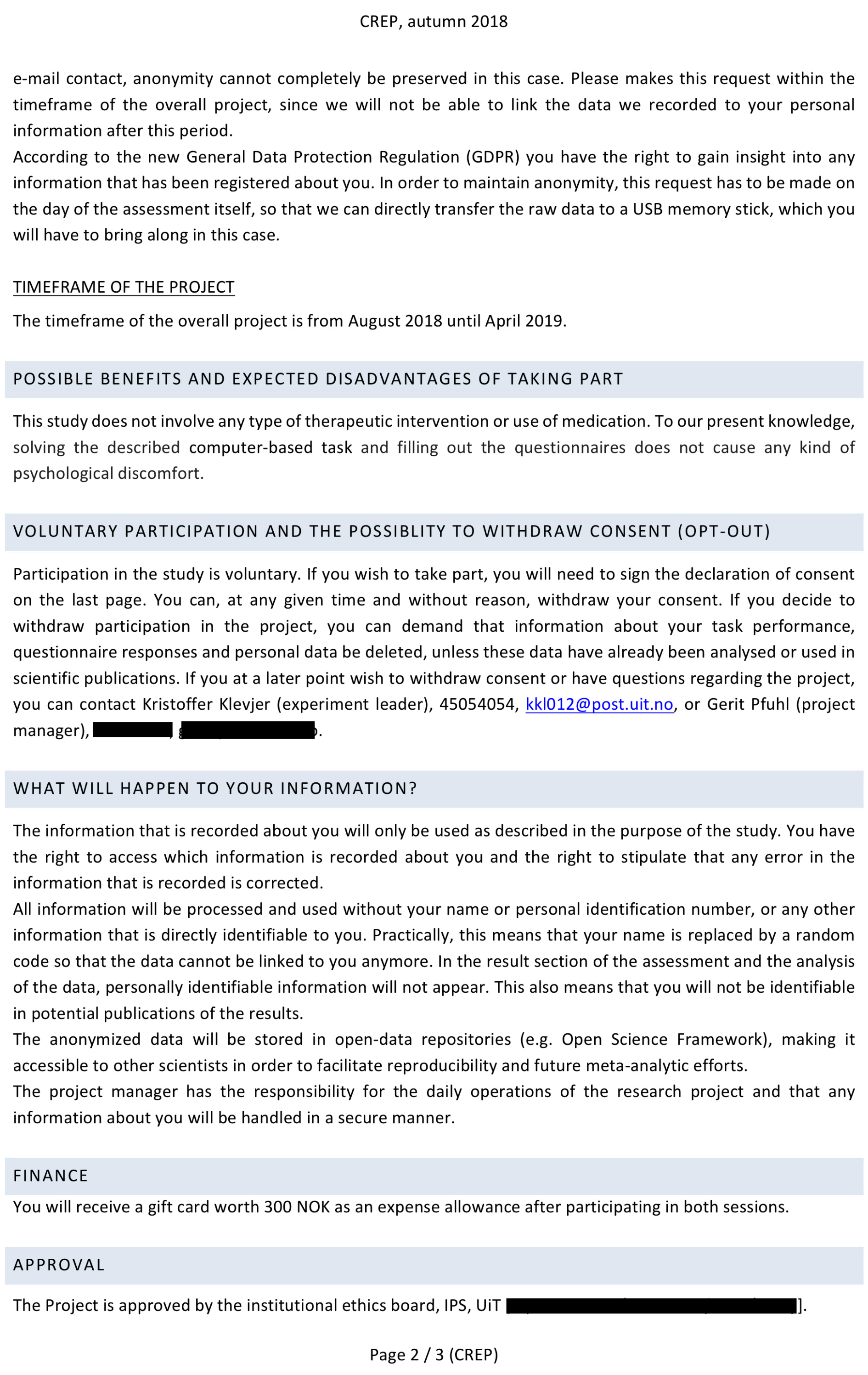
The study consists of two sessions. In session 1, which lasts approximately 55 min you are asked to classify items on a computer, answer a few short questionnaires, and we measure your hand grip. In session 2 you will again classify the items and also perform a working memory task. This session lasts 80 to 90 min.

WHAT IS GOING TO HAPPEN?

You will start with the first session performing the classification task and we measure your hand grip.
After 4 weeks we will test you again on the classification task and now you can also play a working memory task.
In both sessions, you will be asked to fill out a few short questionnaires, which include questions about your demographical background, personal traits and experiences.

WHAT KIND OF INFORMATION WILL WE REGISTER?

The information collected and recorded in this study includes demographics (i.e. age, gender, educational background), responses to the trait- and experience-based questionnaires, and task performance. All data will be anonymized for the analysis.
If you wish feedback on how you did it on the tasks and the questionnaires, you can obtain this information by contacting project manager, Gerit Pfuhl (e-mail: [redacted]). Your anonymized data will then be analyzed by experiment leader Kristoffer Klevjer and forwarded to you through Gerit Pfuhl. Be aware that because of the



e-mail contact, anonymity cannot completely be preserved in this case. Please makes this request within the timeframe of the overall project, since we will not be able to link the data we recorded to your personal information after this period.

According to the new General Data Protection Regulation (GDPR) you have the right to gain insight into any information that has been registered about you. In order to maintain anonymity, this request has to be made on the day of the assessment itself, so that we can directly transfer the raw data to a USB memory stick, which you will have to bring along in this case.

TIMEFRAME OF THE PROJECT

The timeframe of the overall project is from August 2018 until April 2019.

POSSIBLE BENEFITS AND EXPECTED DISADVANTAGES OF TAKING PART

This study does not involve any type of therapeutic intervention or use of medication. To our present knowledge, solving the described computer-based task and filling out the questionnaires does not cause any kind of psychological discomfort.

VOLUNTARY PARTICIPATION AND THE POSSIBLITY TO WITHDRAW CONSENT (OPT-OUT)

Participation in the study is voluntary. If you wish to take part, you will need to sign the declaration of consent on the last page. You can, at any given time and without reason, withdraw your consent. If you decide to withdraw participation in the project, you can demand that information about your task performance, questionnaire responses and personal data be deleted, unless these data have already been analysed or used in scientific publications. If you at a later point wish to withdraw consent or have questions regarding the project, you can contact Kristoffer Klevjer (experiment leader), 45054054, kkl012@post.uit.no, or Gerit Pfuhl (project manager), [redacted] [redacted].

WHAT WILL HAPPEN TO YOUR INFORMATION?

The information that is recorded about you will only be used as described in the purpose of the study. You have the right to access which information is recorded about you and the right to stipulate that any error in the information that is recorded is corrected.

All information will be processed and used without your name or personal identification number, or any other information that is directly identifiable to you. Practically, this means that your name is replaced by a random code so that the data cannot be linked to you anymore. In the result section of the assessment and the analysis of the data, personally identifiable information will not appear. This also means that you will not be identifiable in potential publications of the results.

The anonymized data will be stored in open-data repositories (e.g. Open Science Framework), making it accessible to other scientists in order to facilitate reproducibility and future meta-analytic efforts.

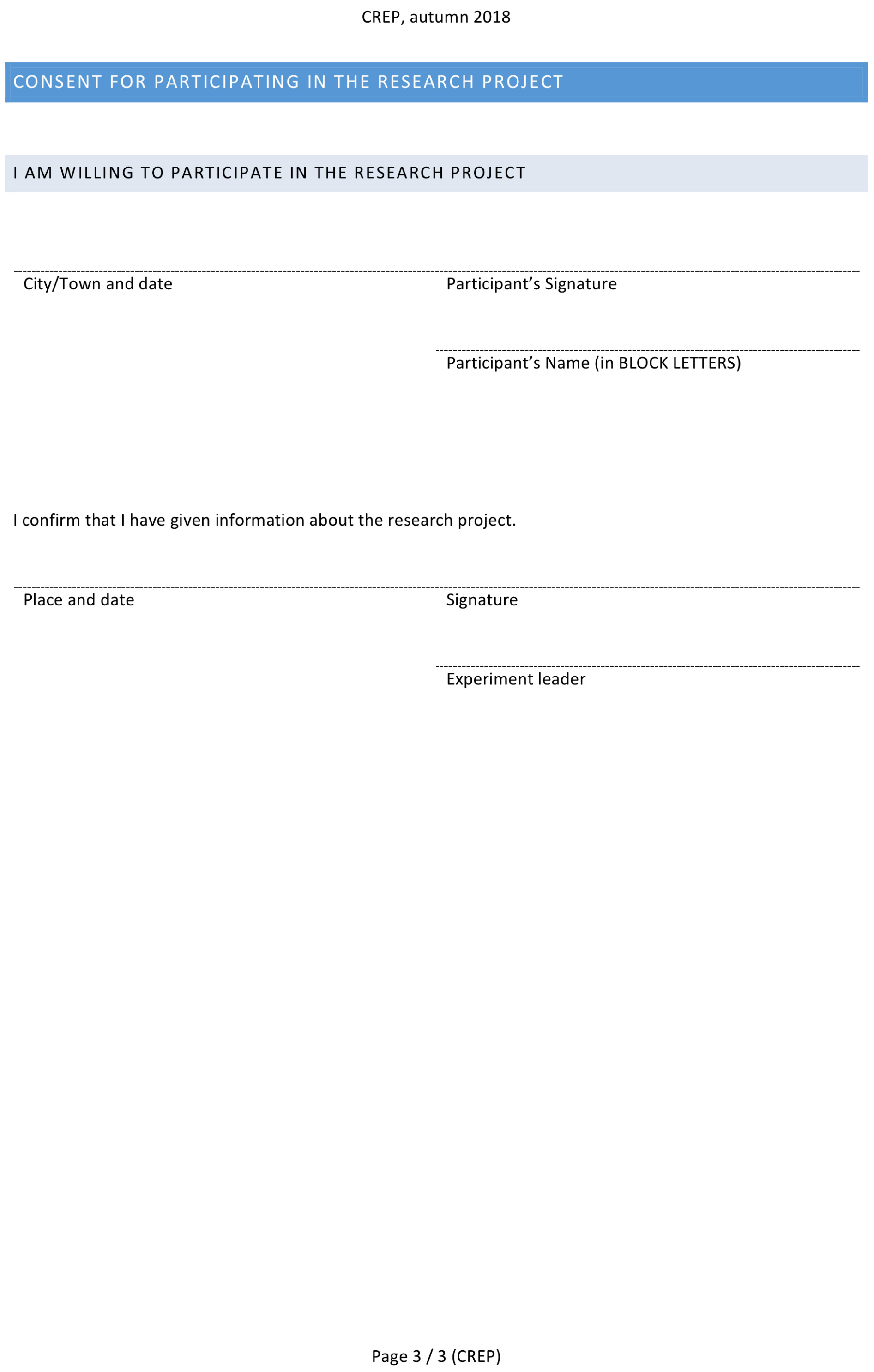
The project manager has the responsibility for the daily operations of the research project and that any information about you will be handled in a secure manner.

FINANCE

You will receive a gift card worth 300 NOK as an expense allowance after participating in both sessions.

APPROVAL

The Project is approved by the institutional ethics board, IPS, UiT [redacted]].



CREP, autumn 2018

CONSENT FOR PARTICIPATING IN THE RESEARCH PROJECT

I AM WILLING TO PARTICIPATE IN THE RESEARCH PROJECT

City/Town and date

Participant’s Signature

Participant’s Name (in BLOCK LETTERS)

I confirm that I have given information about the research project.

Place and date

Signature

Experiment leader

Page 3 / 3 (CREP)

Appendix B

Ethics application (IRB)

Cognitive effort

Application for ethical approval by the Institutional Ethics Committee, IPS, UiT

- **Project title:** Avoidance of cognitive demand: Trait or state – A replication of Kool et al. and reliability assessment.

- **Investigators:**

1. Name: Kristoffer Klevjer
Academic degree: BA
Position: Master student
Workplace: Department of Psychology, UiT – The Arctic University of Norway
Email: kkl012@uit.no
Phone: +47 45 054 054

2. Name: Gerit Pfuhl (PI)
Academic degree: PhD
Position: Associate professor
Workplace: Department of Psychology, UiT – The Arctic University of Norway
Email: [redacted]
Phone [redacted]

- **Expected starting date of the project:** 01.08.2018

- **Expected ending date of the project:** 30.04.2019

- **Are collaborators from other institutions involved in the project?** No

- **Is the project related to other research projects already approved by an ethical committee?** No

- **Is the project part of an education or doctorate?** Yes

- **Does the project involve drug testing?** No

- **Does the project involve collecting new health-related data?** No obvious health information will be obtained.

- **Does the project involve collecting biological material?** No

- **Number of research participants:** 40

- **Recruitment of research participants:** With flyers on the university campus, on social media networks and by personal contact.

- **Will written consent be obtained from all participants?** Yes

- **Inclusion criteria:** Signed informed consent, aged between 18-50 years, no psychiatric/neurological disorder in an acute stage, no drug use within 3 months previous to the assessment (except: tobacco, alcohol), currently no regular intake of central nervous system medications (e.g., antidepressants, antiepileptic drugs, methylphenidate/Ritalin/Concerta), normal or corrected-to-normal eyesight.
- **Exclusion criteria:** Failure to meet the above-mentioned inclusion criteria

- **Describe how participants, the society and/or the scientific community might benefit from the results of the research project.** In 1943, Hull coined the term "law of less work", when all things are equal, people (and animals) tend to show a bias towards the course of action that requires less demand. This has typically been studied using physical demand/effort, although it is often assumed to also include situations involving different cognitive demand as well. Especially evident is this within the field of judgment and decision making / behavioral economics, a field with growing public and scientific interest, and real-life application, where this assumption is often used as an explanation for, or at least a contributor to, various other biases. In 2010, Kool and colleagues set out to test this assumption of a "law of less cognitive demand", and got data in favor of a "avoidance of cognitive demand"-bias. The first part of this project will be a pure replication of that experiment (experiment 3, Decision making and the avoidance of cognitive demand – Kool, McGuire, Rosen and Botvinick, 2010), as a part of a larger international replication project (CREP, https://osf.io/2zw3v/). This is important not only to explore the existence of this bias on its own, but also because of the frequency this assumption is used by the scientific community within this field. And with a growing real-life application of insights from the field, this also have obvious requirements for a solid foundation to be scientifically and ethically sound. Furthermore, we want to investigate to what extent this bias is a trait or state, meaning is it a rather stable character trait to what degree people (possibly) show a bias away from cognitive demand, or does situation or mood a person is in play a role. As a lot of biases is often primarily studied in the lab, insight into the "trait or state"-ness of such an avoidance of cognitive demand bias, could raise the ecological validity, and correct application of knowledge from the field of judgment and decision making and into policies in the real world. Lastly, we will also measure different aspects often included in research into rationality (e.g. need for cognition), as this bias might not be dependent on a maximum processing capacity (often measured by an IQ-test) but rather depend upon other cognitive factors. This is to further explore the characteristics of this bias, to gain insight into such a central assumption in the field, both for the scientific community, and for subsequent use in the real world, both directly and indirectly through other biases that uses this assumption as (part of) it's explanation.

• Describe the potential disadvantages of participating in the research project. What measures will be taken to minimize the impact of these factors?

To our present knowledge, solving the computer-based task and filling out the questionnaires does not cause any kind of psychological discomfort.

In total, each session will last approximately 60 minutes. All participants will be informed about their right to withdraw consent at any time during the experiment without having to give reason for their decision.

• Fees for project manager / co-workers: None

• Compensation for research participants: Non-monetary compensation worth 120 to 500 NOK, depending on how many sessions the participants take part.

• Any conflicts of interest for the project manager/co-workers: None

• Are there restrictions on publication of results of the project? No

• In what form will personally identifiable information and collected data be used and kept?

None of collected data (general data sheet of demographics, task log file, questionnaire responses) will not contain any personally identifiable information, the participants will be instructed on how to generate their own experiment-ID, that they themselves can recreate should they take part in multiple sessions, but at the same time be meaningless to the experimenters, thereby keeping their individual privacy intact and their data anonymous (the first and last letter of their mothers name, the first and last letter of their own last name and the last two digits of their phone number).

• Plan for publishing the results and/or the obtained information and potential further use of results, data, biological material
All data will be stored anonymously. The results obtained within the scope of this project are going to be published in scientific peer-reviewed journals. Anonymized data and analysis scripts will be made publicly available in open-data repositories (e.g. Open Science Framework) to facilitate reproducibility and future meta-analytic efforts.

• Describe the academic and scientific rationale for the selection of data collection:
Behavioral and economic theories have long maintained that actions are chosen so as to minimize demands for exertion or work, a principle sometimes referred to as the law of less work. The data supporting this idea pertain almost entirely to demands for physical effort. However, the same minimization principle has often been assumed also to apply to cognitive demand. Kool et al. (2010) set out to evaluate the validity of this assumption. In 6 behavioral experiments, participants chose freely between courses of action associated with different levels of demand for controlled information processing. Together, the results of these experiments revealed a bias in favor of the less demanding course of action. The bias was obtained across a range of choice settings and demand manipulations and was not wholly attributable to strategic avoidance of errors, minimization of time on task, or maximization of the rate of goal achievement. It is remarkable that the effect also did not depend on awareness of the demand manipulation. Consistent with a motivational account, avoidance of demand displayed sensitivity to task incentives and co-varied with individual differences in the efficacy of executive control. The findings reported, together with convergent neuroscientific evidence, lend support to the idea that anticipated cognitive demand plays a significant role in behavioral decision making. For a finding such as this, a pure replication is important, in order to gain confidence in the existence (or not) of this potential bias. Furthermore, by performing the same task with approximately 4 weeks in between, and compere the individual scores, we can gain insight into whether this seems to be a stable individual characteristic, or if the individual scores do not correlate highly, is more of a situation/state influenced bias.

According to Stanovich, West and Toplak (2016), in their book “The rationality quotient”, rational thinking, like intelligence, is a measurable cognitive competence. Kool et al. (2010) found that the bias seems stronger for people with a high task-switch cost / lower efficacy for executive control, and therefor it seems plausible that measures such as need for cognition (e.g. “the ball and the bat”-item) could be correlated with the degree of expressed avoidance of cognitive demand.

• Summary of the project:
In session 1, the first part of this project will be a pure replication of experiment 3, from Kool et al. (2010). Which is a computerized task, where participants will choose between two different cues, and are then presented with one of two different tasks. Unbeknownst to the participants, the cues differ in their rate of task-switching, and thereby demand for executive control. Then the participants will fill out a short debrief questionnaire to assess the participants’ awareness of the demand manipulation. Next the participants will complete a series of rational thinking skills assessment, including need for cognition, probabilistic-, scientific-, and logical-thinking. They will also do a short physical effort task.

In session 2, the participants will do the effort task and another cognitive effort task, about 4 weeks later. If the individual scores from session 1 and 2, does not correlate highly, this is suggestive of a more state-like nature of the avoidance of cognitive demand bias. A session 3 will be added, which is the same as session 1, but with the inclusion of a small arousal manipulation (performing 5 jumping jacks to increase arousal).

- **Attachments:**

Attachment 1: Consent form (English and Norwegian version)

Appendix C

DST Experimenter script

Greets the participants and introduces myself

- Thank you for being able to come.

- First of all I want you to carefully read this informed consent form.
It informs you of your rights as a research participant.

Hands over the consent form

- Please read it thoroughly, and ask if you have any questions.
If everything looks okay, then please sign at the last page.

- Was everything clear?

Answers any questions, records them and the answers given if it seem at all likely that they might influence the task performance

- Please come with me in here.

Leads the participant into the test room

- If you have a cellphone or any electronic devices
please turn them off

- Now I'll explain what we are going to do here.

- We'll be looking at how you make decisions when solving
tasks, using a computer program.

- All of your data will be kept private, as described in the informed
consent form you just read.

- There is no time limit.

- Now we are ready to begin.

- There will be two coloured circles on the screen, and each time you choose one of them.

- You will then be presented with a number between 1 and 9, with the exception of 5.

- The number can appear in two different colors.
Either blue or orange.

- If the number is in blue, you have to decide if it's higher or lower than 5.

- If the number is above 5, press the right mouse button.
If the number is below 5, press the left mouse button.

- If the number is in orange, you have to decide if it's an odd or even number.

- If the number is even, press the right mouse button.
If the number is odd, press the left mouse button.

- Do you want me to repeat the instructions?

- You'll also get an instruction sheet with these instructions on it.

Answers/repeats the instructions if needed

- Firstly there will be a practice session, where you'll only practice the number judgments and not choose between the circles.

- I will come back after you have finished that session. Please press "0" when you are ready to begin.

I leave the room and wait for the participant, check the practice session score to see that the participant understood the test, otherwise I'll return to the test intro, refer to the instruction sheet, and run the practice session again, as well as record that this happend

- Now we can begin the real experiment, unless you have any other questions?

Answers any questions

- This time you'll be presented with the colored circles, you pick one, and then complete the task the same way you did during the practise session.

- Between each task, you will have to move the mouse cursor to the small white dot in the middle to be able to choose the next circle.

- There is still no time limit. This time you will not receive feedback between each task.

- It's important that you pick from both of the colored circles, but if you start to get a preference for one, you can pick that one more often, if you feel like it.

- It is however important that you don't use simple tactics such as alternating between the circles each trial.

- Rather, you should try in each trial to make a real decision to what circle you want to choose.

- Was everything clear?

Answers any questions, then leave the room and wait for the participant to finish

At this time, the replication is done. The debrief questions will only be presented after test session 2 (as a part of a larger study), please refer to the wiki.

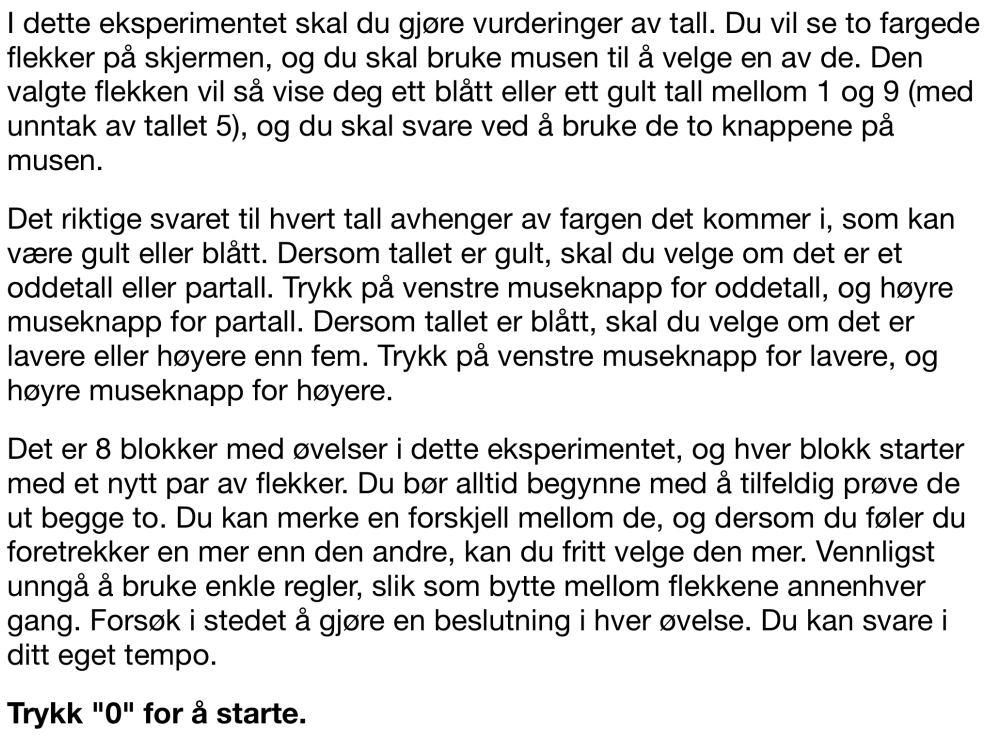
- Then you can come here, fill out this short debrief questionnaire, and then you'll be all done.

- If you know someone else that will be participating in this experiment, please wait till after they are done before discussing it.

- Thanks again for your participation!

Appendix D

DST Handout

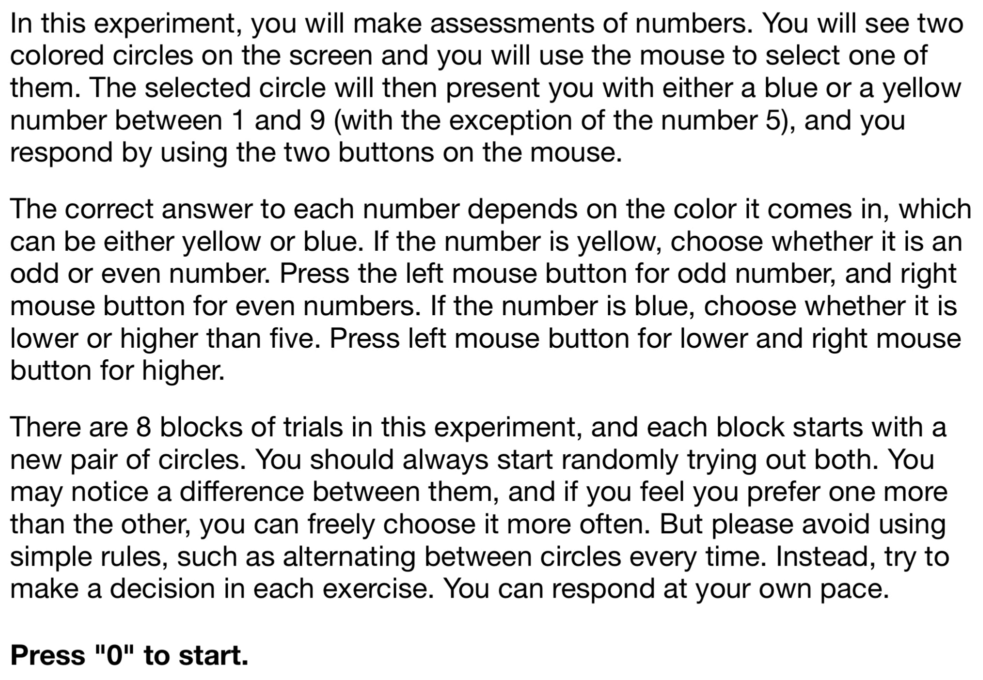


I dette eksperimentet skal du gjøre vurderinger av tall. Du vil se to fargede flekker på skjermen, og du skal bruke musen til å velge en av de. Den valgte flekken vil så vise deg ett blått eller ett gult tall mellom 1 og 9 (med unntak av tallet 5), og du skal svare ved å bruke de to knappene på musen.

Det riktige svaret til hvert tall avhenger av fargen det kommer i, som kan være gult eller blått. Dersom tallet er gult, skal du velge om det er et oddetall eller partall. Trykk på venstre museknapp for oddetall, og høyre museknapp for partall. Dersom tallet er blått, skal du velge om det er lavere eller høyere enn fem. Trykk på venstre museknapp for lavere, og høyre museknapp for høyere.

Det er 8 blokker med øvelser i dette eksperimentet, og hver blokk starter med et nytt par av flekker. Du bør alltid begynne med å tilfeldig prøve de ut begge to. Du kan merke en forskjell mellom de, og dersom du føler du foretrekker en mer enn den andre, kan du fritt velge den mer. Vennligst unngå å bruke enkle regler, slik som bytte mellom flekkene annenhver gang. Forsøk i stedet å gjøre en beslutning i hver øvelse. Du kan svare i ditt eget tempo.

Trykk "0" for å starte.



In this experiment, you will make assessments of numbers. You will see two colored circles on the screen and you will use the mouse to select one of them. The selected circle will then present you with either a blue or a yellow number between 1 and 9 (with the exception of the number 5), and you respond by using the two buttons on the mouse.

The correct answer to each number depends on the color it comes in, which can be either yellow or blue. If the number is yellow, choose whether it is an odd or even number. Press the left mouse button for odd number, and right mouse button for even numbers. If the number is blue, choose whether it is lower or higher than five. Press left mouse button for lower and right mouse button for higher.

There are 8 blocks of trials in this experiment, and each block starts with a new pair of circles. You should always start randomly trying out both. You may notice a difference between them, and if you feel you prefer one more than the other, you can freely choose it more often. But please avoid using simple rules, such as alternating between circles every time. Instead, try to make a decision in each exercise. You can respond at your own pace.

Press "0" to start.

Appendix E

DST Debrief

- Hvordan var det å utføre oppgaven?

- Hvordan valgte du mellom de forskjellige rundingene?

- Utviklet du en preferanse for en av rundingene?

- Var det noe forskjell mellom rundingene?

- For noen av deltakerene hadde den ene av de to rundingene en tendens til å bytte mellom fargene oftere, mens den andre rundingen oftere gjentok den samme fargen. Virket det som om dette var tilfellet for deg?

- Dersom du svarte ja på forrige spørsmål (indikerte at en av rundingene så ut til å bytte mellom fargene hyppigere enn den andre), var dette noe du ble EKSPLISITT klar over UNDER EKSPERIMENTET, eller noe du tenkte på i etterkant?

- What was it like performing the task?

- How did you choose between circles?

- Did you develop a preference for one of the circles?

- Was there any difference between the circles?

- For some participants, one of the two circles had a tendency to switch between colors more often while the other circle tended to repeat the same color. Did it seem like this was the case for you?

- If you answered yes to the previous question (indicating that one of the circle seemed to switch between colors more often), was this something you became EXPLICITLY aware of DURING THE EXPERIMENT, or something that you realized only in retrospect?

Appendix F

RQ-task items

Information in brackets was not provided to the participants, included here to clarify sources and types of items. All the original items were originally in English.

[Introduction]
Du skal nå svare på noen spørsmål og løse noen oppgaver. Svar etter beste evne på spørsmålene og velg det alternativet som passer deg best.
Noen oppgaver vil være vanskelige, andre vil være lettere, gjør ditt beste for å løse dem.

[Item 1 – Heuristic item – 'CRT7' from Toplak, West, and Stanovich, 2011, p. 151]
Simon bestemte seg for å investere 80,000kr i aksjemarkedet en dag tidlig I 2008. Seks måneder etter at han investerte, 17. Juli, hadde aksjene han hadde kjøpt gått ned 50% i verdi. Heldigvis for Simon, fra 17. Juli til 17. Oktober, steg aksjene han hadde kjøpt opp i verdi med 70%. På dette tidspunktet har Simon:

○ Like mye som da han startet (1)

○ Mer enn da han startet (2) [heuristic answer]

○ Tapt penger (3) [correct answer]

[Item 2 – Non-heuristic item – Disjunctive reasoning from Levesque, 1986, p. 85]
Jack ser på Anne, men Anne ser på George. Jack er gift, men George er ikke det. Ser en gift person på en ugift person?

○ Ja (1) [correct answer]

○ Nei (2)

○ Kan ikke fastslås (3) [(possible heuristic answer)]

[Item 3 – Heuristic item – 'Probability matching' from Koehler and James, 2010, p. 669]
I denne oppgaven skal du velge blant 10 par kopper. Hvert par består av 1 blå kopp og 1 gul kopp.
Det er altså 20 kopper totalt, 10 blå kopper og 10 gule kopper.
Det er plasert én femtilapp (50kr) under én av koppene i hvert par.
Måten det ble bestemt hvilken kopp femtilappen ble plassert under var ved å kaste terning.
Terningen har 10 sider, 7 blå sider og 3 gule sider.
Hvis terningen landet på blå er femtilappen under den blå koppen, hvis terningen landet på gul er femtilappen plassert under den gule koppen.

	Velg 1 kopp i hvert par.	
	Blå kopp (1)	Gul kopp (2)

Par 1 (1)	○	○
Par 2 (2)	○	○
Par 3 (3)	○	○
Par 4 (4)	○	○
Par 5 (5)	○	○
Par 6 (6)	○	○
Par 7 (7)	○	○
Par 8 (8)	○	○
Par 9 (9)	○	○
Par 10 (10)	○	○

[10 blue cups: correct answer, 7 blue & 3 yellow: heuristic answer]

[Item 4 – Heuristic item – 'CRT2' from Frederick, 2005, p. 27]
Hvis det tar 5 maskiner 5 minutter å lage 5 leketøy, hvor lang tid tar det for 100 maskiner å lage 100 leketøy? ______Minutter. [5: correct answer, 100: heuristic answer]

[Item 5 – Non-heuristic item – Teigen and Keren, 2007, p. 339]
Se for deg følgende scenario
Fred reiser til jobben med en buss som har avgang en gang i timen. Fred har observert at bussen ankommer før planlagt avgang i 10 % av tilfellene, 0 – 10 minutter etter planlagt avgang i 80% av tilfellene, og den er mer enn 10 minutter forsinket i 10% av tilfellene.

Hvis Fred ankommer busstoppet akkurat i tide og venter i 10 minutter uten at bussen ankommer. Hva er mest sannsynlig? Velg ett svaralternativ.

○ Bussen ankom før tiden (1)

○ Bussen vil fortsatt ankomme (2)

○ Begge deler er like sannsynlig (3) [correct answer]

[Item 6 – Heuristic item – 'CRT3' from Frederick, 2005, p. 27]
I en dam er det et stort område med vannliljer. Hver dag dobler området seg i størrelse. Hvis det tar 48 dager for vannliljene å dekke hele dammen. Hvor lang tid tar det før vannliljene dekker halve dammen? _______Dager. [47: correct answer, 24: heuristic answer]

[Item 7 – Non-heuristic item – G. Gigerenzer, 2007; as cited in Gigerenzer, Gaissmaier, Kurs-Milcke, Schwartz, and Woloshin, 2007, p. 55]
En 50 år gammel kvinne, uten symptomer, deltar i rutinemessig mammografisk screening. Hun tester positivt, er bekymret, og vil vite fra deg om det er helt sikkert at hun har brystkreft eller hva sjansene er. Bortsett fra screeningsresultatene, vet du ingenting annet om denne kvinnen. Hvor mange kvinner som tester positivt har faktisk brystkreft? [small 'natural frequencies' nudge]

- Sannsynligheten for at en kvinne har brystkreft er 1 prosent (prevalens)
- Hvis en kvinne har brystkreft, er sannsynligheten for at hun tester positiv 90 prosent (følsomhet)
- Hvis en kvinne ikke har brystkreft, er sannsynligheten for at hun likevel tester positivt 9 prosent (falsk alarmrate)

Hva er sjansene for at hun har kreft?

○ 9 av 10 (1)

○ 8 av 10 (2)

○ 1 av 10 (3) [correct answer]

○ 1 av 100 (4)

[Item 8 – Heuristic item – 'CRT4' from Toplak, West, and Stanovich, 2011, p. 151; personal correspondence between Toplak, West, and Stanovich, with Frederick, 2011]
Hvis John kan drikke et vannfat (120 liter) på 6 dager, og Mary kan drikke ett vannfat på 12 dager, hvor lang tid vil det ta dem å drikke et vannfat sammen? _______dager.
[4: correct answer, 9: heuristic answer]

[Item 9 – Non-heuristic item – Stanovich, West, and Toplak, 2016, p. 100; built on Stanovich & West, 1998; adapted from Beyth-Marom and Fischoff, 1983]
Tenk deg at du møter David Maxwell. Din oppgave er å vurdere sannsynligheten for at han er universitetsprofessor basert på den informasjonen du vil få. Dette vil bli gjort i to trinn.

Ved hvert trinn vil du få informasjon som du kanskje, eller kanskje ikke, finner nyttig for å gjøre din vurdering. Etter hver bit med informasjon vil du bli bedt om å vurdere sannsynligheten for at David Maxwell er universitetsprofessor. Når du gjør din vurdering må du vurdere all informasjon du har mottatt til det punktet som du anser som relevant.

[(Item 9a)]
Du blir fortalt at David Maxwell deltok på et selskap hvor 25 mannlige universitetsprofessorer og 75 mannlige bedriftsledere deltok, 100 mennesker til sammen.
Spørsmål: Hva tror du sannsynligheten er for at David Maxwell er universitetsprofessor? ___

0 10 20 30 40 50 60 70 80 90 100

Oppgi svaret i prosent (%) ()	

[(Item 9b)]
Du blir fortalt at David Maxwell er medlem av Bjørnens Klubb. 70% av de mannlige universitetsprofessorene ved det tidligere nevnte selskapet var medlemmer av Bjørnens Klubb. 90% av de mannlige bedriftsledere ved selskapet var medlemmer av Bjørnens Klubb.
Spørsmål: Hva tror du sannsynligheten er at David Maxwell er universitetsprofessor? ___

0 10 20 30 40 50 60 70 80 90 100

Oppgi svaret i prosent (%) ()	

[considered correct if 9b < 9a]

[Item 10 – Non-heuristic item – Smullyan, 1978, p. 22; as cited in Toplak and Stanovich, 2011, p. 1285]
Tenk deg at det er tre innbyggere i et fiktivt land, A, B og C, hver av dem er enten en ridder eller en knekt. Riddere forteller alltid sannheten. Knekter lyver alltid.
To personer sies å være av samme type hvis de begge er riddere eller begge er knekter.
A og B gjør følgende uttalelser:
1) A sier at B er en knekt
2) B sier at A og C er av samme type.
Hva er C?

○ Ridder (1)

○ Knekt (2) [correct answer]

○ Kan ikke fastslås (3) [(possible heuristic answer)]

[Item 11 – Heuristic item – 'CRT5' from Toplak, West, and Stanovich, 2011, p. 151; personal correspondence between Toplak, West, and Stanovich, with Frederick, 2011]
Etter en prøve fikk Jerry både den 15. høyeste og 15. laveste skåren i klassen. Hvor mange studenter er det i klassen? _______ studenter. [29: correct answer, 30: heuristic answer]

[Item 12 – Non-heuristic item – Lehman, Lampert, and Nisbett, 1988, p. 442; similar to Wason, 1966]
Du er offentlig helsepersonell på den internasjonale flyplassen i Manila, hovedstaden på Filippinene. En del av din plikt er å kontrollere at alle ankomne passasjerer som ønsker å reise inn i landet (i stedet for bare å bytte fly på flyplassen) har blitt vaksinert mot kolera. Hver passasjer har med seg et helseskort. Én side av kortet angir om passasjeren reiser inn eller bytter fly, og på den andre siden av skjemaet finner du de vaksinene han eller hun har hatt de siste seks månedene.
Hvilke av de følgende kortene vil du trenge å snu for å sjekke? Angi kun de kortene du må sjekke for å være sikker.

Kort 1) Bytter fly.
Kort 2) Innreise.
Kort 3) Vaksinert mot: kolera, hepatitt.
Kort 4) Vaksinert mot: tyfus.

□ Kort 1) Bytter fly (1) [must be un-ticked for correct answer]

□ Kort 2) Innreise (2) [must be ticked for correct answer]

□ Kort 3) Vaksinert mot: kolera, hepatitt (3) [must be un-ticked for correct answer]

□ Kort 4) Vaksinert mot: tyfus (4) [must be ticked for correct answer]

[Item 13 – Non-heuristic item – Toplak, West and Stanovich, 2011, p. 1285]
En lege hadde jobbet med en kur for en mystisk sykdom. Til slutt skapte han et stoff som han mener vil helbrede folk for sykdommen. Før han kan begynne å bruke den regelmessig, må han teste stoffet. Han valgte 300 personer som hadde sykdommen og ga dem stoffet for å se hva som skjedde. Han valgte 100 personer som hadde sykdommen og gav dem ikke stoffet og observerte hva som skjedde. Tabellen nedenfor viser hva resultatet av forsøket var:

	Frisk	
	Ja	Nei
Mottok behandling	200	100
Mottok ikke behandling	75	25

Var dette stoffet positivt eller negativt forbundet med helbredelse for denne sykdommen?

Veldig negativt Veldig positivt

- -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
10

. ()	

[considered correct when answer < 0]

[Item 14 – Heuristic item – 'CRT6' from Toplak, West, and Stanovich, 2011, p. 151; adapted from Dominowski, 1994]
En mann kjøper en gris for 60$, selger den for 70$, kjøper den tilbake for 80$, og selger den til slutt for 90$. Hvor mye har han tjent? _______dollar. [20: correct answer, 10: heuristic answer]

[Introduction]
You will now answer some questions and solve some tasks. Answer to the best of your ability and select the answer that fits you best. Some tasks will be difficult, others will be easier, do your best to solve them.
Thank you in advance!

[Item 1 – Heuristic item – 'CRT7' from Toplak, West, and Stanovich, 2011, p. 151]
Simon decided to invest $8,000 in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point, Simon has:

○ Broken even in the stock market (1)

○ Is ahead of where he began (2) [heuristic answer]

○ Has lost money (3) [correct answer]

[Item 2 – Non-heuristic item – Disjunctive reasoning from Levesque, 1986, p. 85]
Jack is looking at Anne but Anne is looking at George. Jack is married but George is not. Is a married person looking at an unmarried person?

○ Yes (1) [correct answer]

○ No (2)

○ Cannot be determined (3) [(possible heuristic answer)]

[Item 3 – Heuristic item – 'Probability matching' from Koehler and James, 2010, p. 669]
A five dollar bill (5$) is placed under one of the cups in each pair.
The way it was decided which cup the dollar bill should be placed under was by rolling a dice.
The dice has 10 sides, 7 sides are blue and 3 sides are yellow.
If the dice landed on blue the five dollar bill is placed underneath the blue cup, if the dice landed on yellow the five dollar bill is placed underneath the yellow cup.

	Choose 1 cup in each pair	
	Choose 1 cup in each pair (1)	(2)

Pair 1 (1)	○	○
Pair 2 (2)	○	○
Pair 3 (3)	○	○
Pair 4 (4)	○	○
Pair 5 (5)	○	○
Pair 6 (6)	○	○
Pair 7 (7)	○	○
Pair 8 (8)	○	○
Pair 9 (9)	○	○
Pair 10 (10)	○	○

[10 blue cups: correct answer, 7 blue & 3 yellow: heuristic answer]

[Item 4 – Heuristic item – 'CRT2' from Frederick, 2005, p. 27]
CRT2 If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? ______ minutes [5: correct answer, 100: heuristic answer]

[Item 5 – Non-heuristic item – Teigen and Keren, 2007, p. 339]
Consider the following situation:
Fred goes to work by a bus that departs only once every hour. Fred has observed that the bus arrives before schedule in about 10% of the cases, 0–10 min after schedule in 80% of the cases, and is more than 10 min late in 10% of the cases.

Suppose that Fred arrives at the bus stop exactly on time and waits for 10 min without the bus arriving. What is more likely (choose one option):

○ The bus arrived too early (1)

○ The bus will still arrive (2)

○ Both options are equally likely (3) [correct answer]

[Item 6 – Heuristic item – 'CRT3' from Frederick, 2005, p. 27]
In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ____ days [47: correct answer, 24: heuristic answer]

[Item 7 – Non-heuristic item – G. Gigerenzer, 2007; as cited in Gigerenzer, Gaissmaier, Kurs-Milcke, Schwartz, and Woloshin, 2007, p. 55]
A 50-year old woman, no symptoms, participates in routine mammography screening. She tests positive, is alarmed, and wants to know from you whether she has breast cancer for certain or what the chances are. Apart from the screening results, you know nothing else about this woman. How many women who test positive actually have breast cancer? [small 'natural frequencies' nudge]
• The probability that a woman has breast cancer is 1 percent (prevalence)
• If a woman has breast cancer, the probability that she tests positive is 90 percent (sensitivity)
• If a woman does not have breast cancer, the probability that she nevertheless tests positive is 9 percent (false alarm rate)

What are the chances she has cancer?

○ 9 in 10 (1)

○ 8 in 10 (2)

○ 1 in 10 (3) [correct answer]

○ 1 in 100 (4)

[Item 8 – Heuristic item – 'CRT4' from Toplak, West, and Stanovich, 2011, p. 151; personal correspondence between Toplak, West, and Stanovich, with Frederick, 2011]
If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together? _____ days
[4: correct answer, 9: heuristic answer]

[Item 9 – Non-heuristic item – Stanovich, West, and Toplak, 2016, p. 100; built on Stanovich & West, 1998; adapted from Beyth-Marom and Fischoff, 1983]
Imagine yourself meeting David Maxwell. Your task is to assess the probability that he is a university professor based on some information that you will be given. This will be done in two steps. At each

step, you will get some information that you may or may not find useful in making your assessment. After each piece of information you will be asked to assess the probability that David Maxwell is a university professor. In doing so, consider all the information you have received to that point if you consider it to be relevant.

[(Item 9a)]
You are told that David Maxwell attended a party in which 25 male university professors and 75 male business executives took part, 100 people all together.
Question: What do you think the probability is that David Maxwell is a university professor? ___

	0 10 20 30 40 50 60 70 80 90 100
Answer in percentage (%) ()	

[(Item 9b)]
You are told that David Maxwell is a member of the Bear's Club. 70% of the male university professors at the above mentioned party were members of the Bear's Club. 90% of the male business executives at the party were members of the Bear's Club.
Question: What do you think the probability is that David Maxwell is a university professor?

	0 10 20 30 40 50 60 70 80 90 100
Answer in percentage (%) ()	

[considered correct if 9b < 9a]

[Item 10 – Non-heuristic item – Smullyan, 1978, p. 22; as cited in Toplak and Stanovich, 2011, p. 1285]
Imagine that there are three inhabitants of a fictitious country, A, B, and C, each of whom is either a knight or a knave. Knights always tell the truth. Knaves always lie.
Two people are said to be of the same type if they are both knights or both knaves.
A and B make the following statements:
1) A says that B is a knave, 2) B says that A and C are of the same type. What is C?

○ Knight (1)

○ Knave (2) [correct answer]

○ Cannot be determined (3) [(possible heuristic answer)]

[Item 11 – Heuristic item – 'CRT5' from Toplak, West, and Stanovich, 2011, p. 151; personal correspondence between Toplak, West, and Stanovich, with Frederick, 2011]
Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class? _______ students [29: correct answer, 30: heuristic answer]

[Item 12 – Non-heuristic item – Lehman, Lampert, and Nisbett, 1988, p. 442; similar to Wason, 1966]
You are a public health official at the international airport in Manila, capital of the Philippines. Part of your duty is to check that every arriving passenger who wishes to enter the country (rather than just change planes at the airport) has had an inoculation against cholera. Every passenger carries a health form. One side of the form indicates whether the passenger is entering or in transit, and the other side of the form lists the inoculations he or she has had in the past six months.
Which of the following forms would you need to turn over to check? Indicate only those forms you would have to check to be sure.
Box 1)Transit.
Box 2) Entering.
Box 3) Inoculated against: cholera, hepatitis.
Box4) Inoculated against: typhoid.

□ Box 1 (1) [must be un-ticked for correct answer]

□ Box 2 (2) [must be ticked for correct answer]

□ Box 3 (3) [must be un-ticked for correct answer]

□ Box 4 (4) [must be ticked for correct answer]

[Item 13 – Non-heuristic item – Toplak, West and Stanovich, 2011, p. 1285]
A doctor had been working on a cure for a mysterious disease. Finally, he created a drug that he thinks will cure people of the disease. Before he can begin to use it regularly, he has to test the drug. He selected 300 people who had the disease and gave them the drug to see what happened. He selected 100 people who had the disease and did not give them the drug in order to see what happened. The table below indicates what the outcome of the experiment was:

	Cured	
	Yes	No
Treatment present	200	100
Treatment absent	75	25

Is this treatment positively or negatively associated with the cure for this disease?

Strong negative association	Strong positive association

	- -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 10
()	

[considered correct when answer < 0]

[Item 14 – Heuristic item – 'CRT6' from Toplak, West, and Stanovich, 2011, p. 151; adapted from Dominowski, 1994]

A man buys a pig for $60, sells it for $70, buys it back for $80, and sells it finally for $90. How much has he made? _____ dollars [20: correct answer, 10: heuristic answer]

Appendix G

Need for Cognition items

As seen in Cacioppo, Petty and Kao, 1984, p. 306-307 (English version)

	Passer svært dårlig (1)	(2)	(3)	(4)	Passer svært bra (5)
Jeg foretrekker komplekse fremfor enkle oppgaver/problemer. (1)	○	○	○	○	○
Jeg liker å ha ansvar for situasjoner som krever mye tenkning (2)	○	○	○	○	○
Tankevirksomhet er ikke det jeg synes er mest gøy. (3)	○	○	○	○	○
Jeg gjør heller noe som krever lite tankearbeid, fremfor noe som utfordrer min tankekapasitet (evne). (4)	○	○	○	○	○
Jeg prøver å forutse og unngå situasjoner hvor det er en sjanse for at jeg må tenke grundig/i dybden om noe. (5)	○	○	○	○	○
Jeg finner det tilfredsstillende å fundere og "gruble" lenge og grundig på problemer/ oppgaver jeg kan løse. (6)	○	○	○	○	○
Jeg tenker bare så "hardt" og grundig som det kreves i situasjonen. (7)	○	○	○	○	○
Jeg foretrekker å tenke på mindre, daglige prosjekter fremfor oppgaver/ prosjekter som tar tid. (8)	○	○	○	○	○
Jeg liker oppgaver som krever lite tankearbeid når en først har lært det. (9)	○	○	○	○	○
Ideen om å bruke min intellektuelle kapasitet til å komme meg til topps virker fristende for meg. (10)	○	○	○	○	○
Jeg setter stor pris på oppgaver som går ut på å finne nye løsninger på problemer. (11)	○	○	○	○	○

Å lære seg nye måter å tenke på fascinerer ikke meg i særlig grad. (12)	○	○	○	○	○
Jeg foretrekker at livet mitt er fylt med oppgaver og "puzzles" som jeg må løse. (13)	○	○	○	○	○
Abstrakt tenking appellerer til meg. (14)	○	○	○	○	○
Jeg foretrekker en oppgave som er intellektuell, vanskelig og viktig, fremfor en som i noen grad er viktig, men som ikke krever mye tankearbeid. (15)	○	○	○	○	○
Jeg føler lettelse, mer enn tilfredsstillelse, etter jeg har løst en oppgave som krever mye mental kapasitet/innsats. (16)	○	○	○	○	○
For meg er det nok at noe fører til at jobben blir gjort, jeg bryr meg ikke om hvordan og hvorfor det virker. (17)	○	○	○	○	○
Jeg ender ofte opp med å fundere og gruble over forhold, selv om de ikke får noen innflytelse på meg personlig. (18)	○	○	○	○	○

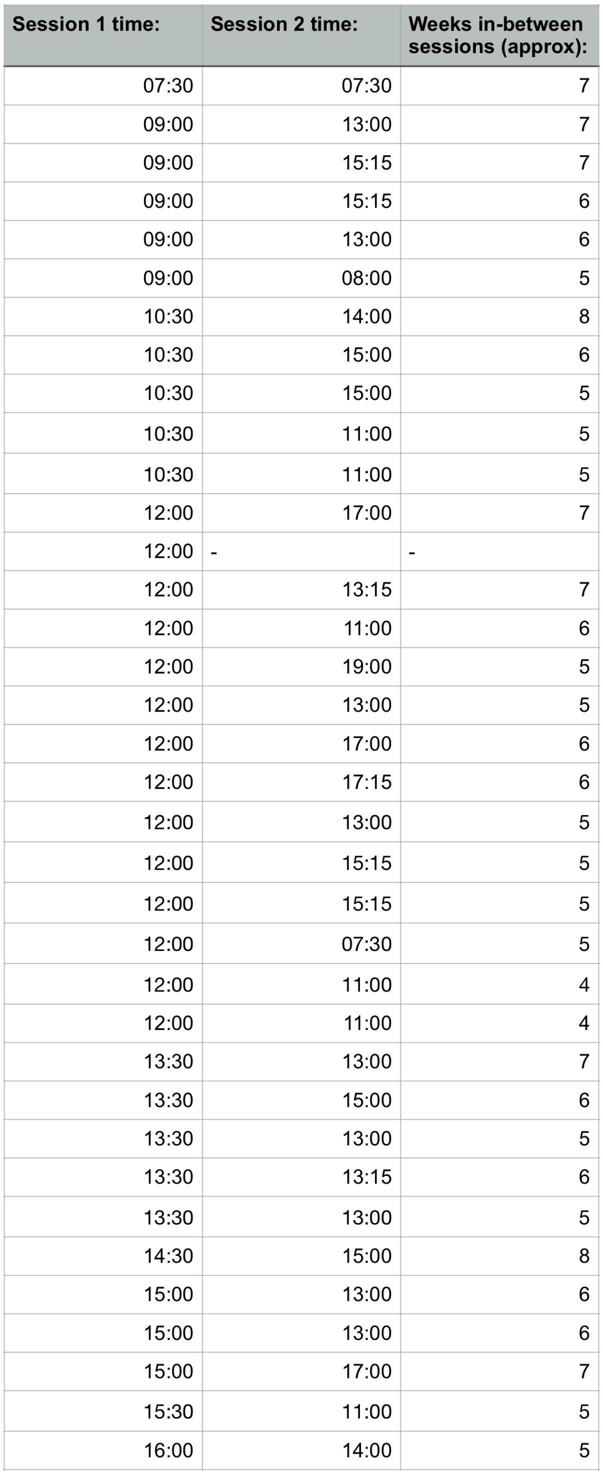
	Very uncharacteristic of me (1)	(2)	(3)	(4)	Very characteristic of me (5)
I prefer complex to simple problems. (1)	○	○	○	○	○
I like to have the responsibility of handling a situation that requires a lot of thinking. (2)	○	○	○	○	○
Thinking is not my idea of fun. (3)	○	○	○	○	○
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities. (4)	○	○	○	○	○
I try to anticipate and avoid situations where there is a likely chance I will have to think in depth about something. (5)	○	○	○	○	○
I find satisfaction in deliberating hard and for long hours. (6)	○	○	○	○	○
I only think as hard as I have to. (7)	○	○	○	○	○
I prefer to think about small daily projects to long term ones. (8)	○	○	○	○	○
I like tasks that require little thought once I’ve learned them. (9)	○	○	○	○	○
The idea of relying on thought to make my way to the top appeals to me. (10)	○	○	○	○	○
I really enjoy a task that involves coming up with new solutions to problems. (11)	○	○	○	○	○

Learning new ways to think doesn't excite me very much. (12)	○	○	○	○	○
I prefer my life to be filled with puzzles I must solve. (13)	○	○	○	○	○
The notion of thinking abstractly is appealing to me. (14)	○	○	○	○	○
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought. (15)	○	○	○	○	○
I feel relief rather than satisfaction after completing a task that requires a lot of mental effort. (16)	○	○	○	○	○
It's enough for me that something gets the job done; I don't care how or why it works. (17)	○	○	○	○	○
I usually end up deliberating about issues even when they do not affect me personally. (18)	○	○	○	○	○

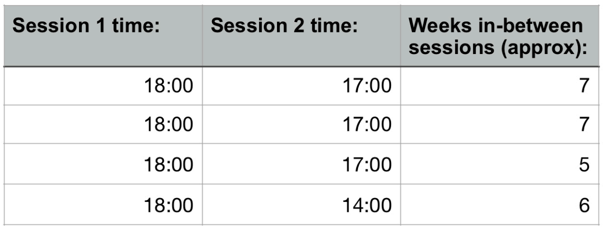
Appendix H

Session dates and times

Only approximate time between sessions provided, in order to protect participants anonymity.



Session 1 time:	Session 2 time:	Weeks in-between sessions (approx):
07:30	07:30	7
09:00	13:00	7
09:00	15:15	7
09:00	15:15	6
09:00	13:00	6
09:00	08:00	5
10:30	14:00	8
10:30	15:00	6
10:30	15:00	5
10:30	11:00	5
10:30	11:00	5
12:00	17:00	7
12:00	-	-
12:00	13:15	7
12:00	11:00	6
12:00	19:00	5
12:00	13:00	5
12:00	17:00	6
12:00	17:15	6
12:00	13:00	5
12:00	15:15	5
12:00	15:15	5
12:00	07:30	5
12:00	11:00	4
12:00	11:00	4
13:30	13:00	7
13:30	15:00	6
13:30	13:00	5
13:30	13:15	6
13:30	13:00	5
14:30	15:00	8
15:00	13:00	6
15:00	13:00	6
15:00	17:00	7
15:30	11:00	5
16:00	14:00	5



Session 1 time:	Session 2 time:	Weeks in-between sessions (approx):
18:00	17:00	7
18:00	17:00	7
18:00	17:00	5
18:00	14:00	6