UiT The Arctic University of Norway

Faculty of Health Sciences,

The role of cognitive effort in decision-making and reasoning errors

Reasoning errors: Beyond insufficient effort - implicating the Locus Coeruleus - Norepinephrine system

Martin Jensen Mækelæ

A dissertation for the degree of Philosophiae Doctor - February 2024

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to thank my main supervisor Gerit Pfuhl. Thank you for your guidance, supervision, and mentoring. I truly appreciate how generous you have been with your time. Thank you for sharing your experience, knowledge, and wisdom. For believing in me, challenging me, and for allowing me to pursue my interests. Without years of collaboration this thesis would not be possible.

Secondly, I would like to thank UiT - The Arctic University of Norway, and IPS – The Institute of Psychology. To my professors, supervisors, classmates, colleagues, and friends, thank you for nearly a decade of learning, hard work and companionship. My time at UiT will remain an important and treasured part of my life story.

I would like to thank all the research participants who have shared their time and effort for the advancement of science. This work would not be possible without your participation.

Lastly, I would like to thank my family, friends, and partner. Thank you for your love and unwavering support. Thank you for always believing in me, supporting me, and sharing your time, thoughts, love and laughter.

Abbreviations

AOI	Areas of Interest
BR	Base-rate task
COG-ED	Cognitive effort discounting paradigm
DST	Demand selection task
LC	Locus Coeruleus
LC-NE	Locus Coeruleus – Norepinephrine
NFC	Need for cognition scale
NE	Norepinephrine
N-TLX	NASA task load index
RQ	Rational reasoning battery
UiT	UiT–The Arctic University of Norway
WashU	Washington University in St. Louis

List of papers

Paper 1

Mækelæ, M. J., Klevjer, K., Westbrook, A., Eby, N. S., Eriksen, R., & Pfuhl, G. (2023). Is it cognitive effort you measure? Comparing three task paradigms to the Need for Cognition scale. *Plos one*, 18(8), e0290177. https://doi.org/10.1371/journal.pone.0290177

Paper 2

Mækelæ, M. J., Kreis, I. V., & Pfuhl, G. (2024). Teleological reasoning bias is predicted by pupil dynamics: Evidence for the extensive integration account of bias in reasoning. *Psychophysiology*, e14532. https://doi.org/10.1111/psyp.14532

Paper 3

Mækelæ, M. J., Kreis, I., & Pfuhl, G. (2024). The influence of visual attention and cognitive effort on base-rate neglect. Manuscript not submitted.

Summary

Cognitive effort is highly familiar in everyday life and may influence our decisions and task performance. However, researchers have struggled to both define and measure cognitive effort. A range of tools measuring cognitive effort has been developed within different lines of research. Yet it is unclear to what degree these tools are related and if they are measuring the same cognitive effort construct. Furthermore, the influential default-interventionist dual-process account proposes that a lack of cognitive effort is a significant source of errors in reasoning and decision-making. However, an accumulating body of research contradicts predictions from the default-interventionist account, giving rise to a new generation of dual-process models. Notably, a separate line of research applying single-process sequential sampling models proposes that bias in decision-making is exacerbated by more extensive integration of evidence. These separate lines of research make opposing predictions regarding pupil dilation. Pupil size can be used as an indicator of both cognitive effort and Locus Coeruleus – Norepinephrine activity. The default-interventionist account predicts that errors in reasoning should be associated with smaller pupil dilations, I.e. less cognitive effort. The extensive integration account proposes that larger pupil dilations, indicating low levels of norepinephrine and neural gain, leads to more extensive integration and more bias in reasoning. Thus, competing models and frameworks with opposing predictions regarding cognitive effort and errors in decision-making can be tested by measuring pupil size during performance on reasoning tasks.

The aims of the thesis were to investigate the role of cognitive effort in decision-making and errors in reasoning. Further, to evaluate tools measuring cognitive effort in decision-making, and lastly to evaluate competing dual-process models and alternative frameworks of decision-making. Paper 1 assessed the shared variance between three behavioral measures of cognitive effort and their relationship to the need for cognition scale. Additionally, working memory capacity and subjective mental effort of the task paradigms was measured. The results showed no relation between the three behavioral measures of cognitive effort. However, two of the measures were related to need for cognition and working memory capacity. Contrary to dual-process model predictions, performance on a battery of rational reasoning tasks was negatively related to subjective mental effort on the tasks. Indicating that more cognitive effort was associated with errors in reasoning. Paper 2 and Paper 3 applied pupillometry to assess cognitive effort during decision-making tasks. Paper 2 found that larger pupil dilations, indicating more cognitive effort, was associated with more errors on a teleological reasoning task, thus finding support for the extensive integration account of bias in reasoning. Paper 3 measured eye-gaze and pupil dilation in two separate versions of a base-rate task. The results were partly mixed. However, evidence suggested base-rate neglect was a significant source of bias on the task. Further, larger pupil dilations associated with conflict detection and cognitive decoupling were related to correct responses. Implicating the Locus Coeruleus – Norepinephrine system in conflict detection and overriding of erroneous responses. The thesis concludes that a lack of cognitive effort is not a general cause of decision-making errors. Rather, errors in reasoning can be associated with both more and less cognitive effort, dependent on the task. Researchers should be mindful of the tools available when measuring cognitive effort as tools differ in reliability, validity, and may measure different aspects of cognitive effort. The results from the thesis largely oppose dual-process models.

The results are in line with sequential sampling models and the extensive integration account of bias in reasoning, highlighting the role of Locus Coeruleus in decision making and reasoning errors. Notably, sequential sampling models were not tested to the same extent as dual-process models in this thesis and the results should be considered preliminary. However, the tools and methodology applied in the thesis may suggest a path forward for future research on errors in decision-making.

1 Introduction

The phenomenal feeling of cognitive effort is highly familiar to everyday life. Cognitive effort may influence both the decision to engage in a task and determine the successful completion of the task. However, researchers have struggled to both define what cognitive effort is and how to measure cognitive effort (Shenhav, Musslick, et al., 2017; Thomson & Oppenheimer, 2022; Westbrook & Braver, 2015). The ubiquitous nature of cognitive effort remains part of the challenge as nearly all task performance is dependent on both cognitive ability and the willingness to perform the task at hand, that is, expending the required cognitive effort. Despite the challenge of both defining cognitive effort and disentangling effort from performance and ability, several researchers have noted that a fundamental property of cognitive effort is the tendency to minimize cognitive effort, all else being equal (Allport, 1954; Hull, 1943; Kool et al., 2010; Solomon, 1948; Zipf, 1949). This is known as the "law of least effort", which have been applied to effort in both the physical and cognitive domain (Allport, 1954; Hull, 1943; Zipf, 1949). In the last two decades the study of cognitive effort has revealed that cognitive effort is treated as costly(Westbrook et al., 2013), although, the nature of this cost is uncertain. Cognitive effort seems to be expended in relation to the expected attainment of goals and rewards and may be subject to multiple trade-offs (Aston-Jones & Cohen, 2005; Kool & Botvinick, 2014; Kurzban et al., 2013; Shenhav et al., 2013). Cognitive effort is generally experienced as aversive, this to the point where humans will both forego rewards and endure pain to avoid cognitive effort (Vogel et al., 2020; Westbrook et al., 2013). However, some individuals may be more inclined to expend, and value cognitive effort (Cacioppo et al., 1996; Cacioppo & Petty, 1982; Inzlicht et al., 2018). Of particular importance to the understanding of cognitive effort, and the influence of cognitive effort, is the relationship between cognitive effort and decision-making. Firstly, the investigation of cognitive effort through decision-making has significantly advanced research on cognitive effort (Shenhav, Musslick, et al., 2017; Westbrook et al., 2021). In particular, the development of task paradigms for measuring cognitive effort costs and cognitive effort avoidance through decision-making have been fundamental to this advancement (Kool et al., 2010; Westbrook et al., 2013). These task paradigms have advanced our knowledge of both the neural mechanisms and computations involved in tracking cognitive effort costs and the allocation of cognitive effort (Sayalı & Badre, 2021; Shenhav, Musslick, et al., 2017; Westbrook et al., 2020, 2021). However, the relationship between these newly developed task paradigms has until now not yet been properly investigated (Thomson & Oppenheimer, 2022). Thus, it has been unknown to what extent these task paradigms are measuring the same cognitive effort construct. Second, the tendency to minimize cognitive effort has been proposed as one of the main reasons for human errors in reasoning and decision-making (J. St. B. T. Evans, 2006; Kahneman, 2011; Shah & Oppenheimer, 2008; Stanovich, 2009a). This proposal has been advanced through research on heuristics and biases and dual-process models of reasoning and decision-making (J. St. B. T. Evans, 2006; Kahneman, 2011). The classical default-interventionist dual-process account proposes that errors in reasoning are largely a result of fast effortless processing and proposes that slower more effortful deliberate processing generally leads to better decision-making and fewer errors in reasoning (J. St. B. T. Evans, 2006; J. St. B. T. Evans & Stanovich, 2013; Kahneman, 2011; Stanovich, 2009a). This notion of errors in reasoning resulting from a lack of cognitive effort has been so fundamental that researchers have used performance on heuristics and bias tasks as measures of cognitive effort, often labeled as measures of deliberate (reflective) thinking or an analytic thinking style (Frederick, 2005; Pennycook, Cheyne, et al., 2012; Pennycook et al., 2015a; Shenhav, Rand, et al., 2017; Stanovich, 2016; Trippas et al., 2015). However, an accumulating body of research is questioning both assumptions of classical dual-process models and the relationship between cognitive effort and reasoning errors. (Bago & De Neys, 2017; Newman et al., 2017; Raoelison et al., 2020; Thompson et al., 2018). This has given rise to a new generation of dual-process models which remain to be properly tested (De Neys & Pennycook, 2019; Pennycook et al., 2015b; Thompson et al., 2018). A novel approach to study dual-process models and cognitive effort in reasoning and decision-making is through the use of pupillometry, as the pupil is known to track cognitive effort (Beatty & Lucero-Wagoner, 2000; Kahneman, 1973; Mathot, 2018; van der Wel & van Steenbergen, 2018). Recently, Eldar et al. (2021) applied pupillometry to test assumptions of the default-interventionist dual-process model against an alternative framework of decision-making, namely sequential sampling models (Boag et al., 2023; N. J. Evans & Wagenmakers, 2020; B. U. Forstmann et al., 2016; Ratcliff et al., 2016, p. 20) and finding evidence in favor of sequential sampling models of decision-making. Importantly, in addition to reflecting cognitive effort, the pupil is closely linked with the Locus Coeruleus (LC) – Norepinephrine (NE) system. The LC is a brainstem nucleus with far-reaching connections and is known to regulate the sleep-wake cycle, arousal, and attention, in addition to influencing several more specific cognitive functions (Berridge & Waterhouse, 2003; Bouret & Sara, 2005; Chandler et al., 2014; McBurney-Lin et al., 2019; McGaughy et al., 2008; Poe et al., 2020; Spencer & Berridge, 2019; Takeuchi et al., 2016; Viglione et al., 2023; Waterhouse & Navarra, 2019). Importantly, the LC-NE system influences brain wide changes in neural gain (Eldar et al., 2013), which have been proposed to influence evidence weighting in sequential sampling models of decision-making (Eldar, Cohen, et al., 2016; Eldar et al., 2021).

The overarching aim of this thesis was to assess the role of cognitive effort in decision-making and errors in reasoning. To achieve this, we evaluated tools measuring cognitive effort in decision-making, and tested competing dual-process models and alternative frameworks of decision-making. In the following introduction, I will first define and present theories of cognitive effort. Second, I will review dual-process models of decision-making and theoretical developments. Third, I will introduce pupillometry as a tool for investigating cognitive effort in decision-making. Fourth, I will introduce the LC-NE system. Fifth, I will briefly introduce sequential sampling models as an alternative framework for decision-making. Lastly, I will expand upon the aims of the thesis.

1.1 Cognitive effort and decision making

The lack of a common definition of cognitive effort has been a challenge for both conducting research on cognitive effort and integration of research across fields, this has been described as the “effort problem” (Thomson & Oppenheimer, 2022). To describe what is meant by cognitive effort here, we refer to two complementary earlier descriptions.

*“*At a coarse level, “effort” refers to the degree of engagement with demanding tasks.” (Westbrook & Braver, 2015, p. 3).

> Effort is what mediates between (a) the characteristics of a target task and the subject’s available information-processing capacity and (b) the fidelity of the information-processing operations actually performed, as reflected in task performance. The first two factors, task characteristics and capacity, determine what level of performance is attainable in principle. Effort refers to the set of intervening processes that determine what level of performance will in fact be realized. (Shenhav, Musslick, et al., 2017, pp. 100 - 101)

Accordingly, cognitive effort refers to the degree of engagement with a task or the cognitive resources employed for task performance. However, cognitive effort will always depend on both the characteristics of the task (task demands), and the cognitive resources available (cognitive ability). Thus, cognitive effort represents the mobilization of cognitive processing resources leading to a certain level of performance being achieved, dependent on both task characteristics and cognitive ability.

Cognitive effort needs to be distinguished from mental operations which execution requires cognitive effort. Although it is not clear why some mental operations require more effort than others, a common factor that seems to require cognitive effort is top-down cognitive control (Botvinick & Braver, 2015; Botvinick & Cohen, 2014; McGuire & Botvinick, 2010; Musslick et al., 2018; Sayalı & Badre, 2019; Schneider & Shiffrin, 1977; Shenhav et al., 2013; Shenhav, Musslick, et al., 2017). Information processing may be viewed as a continuum of automatization (Schneider & Shiffrin, 1977). On one end of the continuum there are processes which have been highly automatized (often through repeated practice) and require little to no effort and impose little interference on other processes. On the other end of the continuum, there are processes which rely on control demanding resources. Cognitive control allows for more flexible information processing, allow for contextual influences on information processing, and can reorganize information processing away from the default. Cognitive control is a cognitively effortful process and heavily interferes with other processes. Note, the end points of this continuum may also be seen as separate processes (see 1.2 Dual-process models of decision-making). Examples of effortful cognitive (executive) control functions are working memory maintenance and sustained attention, updating, task switching, inhibition and overriding of habitual responses (Botvinick & Braver, 2015; Braver, 2012; Friedman & Miyake, 2017; Kahneman, 1973; E. K. Miller & Cohen, 2001; Miyake et al., 2000; Monsell, 2003). Thus, cognitive effort is not the same as cognitive control, but cognitive effort may be exerted through cognitive control. Accordingly, cognitive effort is not a unidimensional construct and may refer to several related phenomena such as, the phenomenal feeling of cognitive effort, the allocation of cognitive resources (deciding to expend cognitive effort), and the execution of demanding cognitive operations.

The ubiquity of effort minimization as a fundamental principle or “law” of behavior was early noted by several theorists, referring to the tendency for humans and animals to minimize the total amount of effort, all else being equal (Allport, 1954; Hull, 1943; Solomon, 1948; Zipf, 1949). Later, self-report measures of experienced cognitive effort (Hart & Staveland, 1988) and of enjoyment of cognitive effort was created (Cacioppo & Petty, 1982). The need for cognition (NFC) questionnaire measuring the tendency to engage in and enjoy effortful cognitive activity has proven to be a reliable and stable trait measure of individual differences in cognitive effort (Cacioppo & Petty, 1982; Hussey & Hughes, 2020). Research on cognitive effort through self-reported NFC, with questions such as "I find satisfaction in deliberating hard and for long hours", has elucidated characteristics of individuals with different dispositions for engaging in demanding mental work (Cacioppo et al., 1996). Showing that cognitive effort is related to task performance and cognitive ability, in addition to important outcomes such as academic achievement and attitude formation (Cacioppo et al., 1996; Colling et al., 2022). However, self-report measures may be subject to bias and there are concerns regarding the validity of self-report measures (Paulhus & Vazire, 2007). Furthermore, to understand cognitive effort, in addition to trait measures of cognitive effort such as NFC, it is important to investigate state dependent changes and actualized cognitive effort expenditure, performance related changes, and the cognitive systems as well as the computational and physiological underpinnings of cognitive effort (Shenhav, Musslick, et al., 2017; Thomson & Oppenheimer, 2022; Westbrook & Braver, 2015).

In recent years, cognitive effort has been framed in economic terms and studied through decision-making paradigms (Kool et al., 2010; Shenhav et al., 2013; Shenhav, Musslick, et al., 2017; Westbrook et al., 2013; Westbrook & Braver, 2015). Experimental evidence of cognitive demand avoidance was presented by Kool et al. (2010) by the development of the demand selection task (DST). In the DST participants were asked to make a series of choices between two similar visual stimuli. When a choice was made participants were presented with a series of numbers. Depending on the color of the number participants either had to make a parity or a magnitude judgement. By leveraging the fact that task switching requires cognitive control and cognitive effort, the only difference between the two visual stimuli were the number of task switches (90% vs 10%). Across multiple experiments it was shown that most people tend to avoid the more mentally demanding stimulus, all else being equal. Studies using variants of the DST have shown that humans learn to avoid mental effort adaptively while tracking effort costs and cost-prediction errors (Nagase et al., 2018). Additionally, brain imaging studies have shown decreased activity in reward regions due to increasing cognitive demand (Botvinick et al., 2009). Further, Westbrook et al. (2013) were able to quantify the cost of cognitive effort and showed individual differences in cognitive effort costs by developing the cognitive effort discounting paradigm (COG-ED). The COG-ED relies on a working memory task, the n-back task (Owen et al., 2005), where mental demand can be manipulated by how many pieces of information must be continually updated in working memory. After experiencing different demand levels participants are asked to make a series of choices where they have to decide if they want to complete a harder task for more money or an easier task for less money. Demand and offer amounts are manipulated and titrated to find a subjective indifference point between options (Westbrook et al., 2013). Thus, one can calculate the subjective value of mental work across load levels for each individual. This task paradigm has revealed stable monetary discounting due to cognitive effort costs (Westbrook et al., 2013, 2019, 2020). The cost of cognitive effort may be partly attributed to the phenomenal feeling of cognitive effort, which is usually described as aversive. A study relying on a variant of COG-ED showed that negative experience of cognitive effort is so strong that some individuals are willing to endure pain to avoid it (Vogel et al., 2020). However, it should be noted that cognitive effort may also be valued, learned to be associated with rewards, and followingly become rewarding in itself (Clay et al., 2022; Eisenberger, 1992; Inzlicht et al., 2018).

The development of task paradigms investigating cognitive effort through decision-making, such as the DST and COG-ED, in addition to other task paradigms (Apps et al., 2015; Chong et al., 2017; Collins et al., 2017; Fallon et al., 2017; Froböse et al., 2018; Massar et al., 2015; Sidarus et al., 2019; Vassena et al., 2014), have allowed for greater insights into the neural mechanisms involved in cognitive effort. In a neuro-economic framework cognitive effort is usually treated as carrying a disutility that is expended in relation to expected rewards, goals or leisure (Kool & Botvinick, 2014, 2018; Shenhav et al., 2013; Shenhav, Musslick, et al., 2017; Westbrook et al., 2013). The subjective value of cognitive effort seems to be encoded by a domain-general valuation network centered in the ventromedial (vm) pre-frontal cortex (PFC) (Westbrook et al., 2019). It has been proposed that the anterior cingulate cortex (ACC) integrates information on potential rewards, the cost of cognitive control, task demands and performance, in order to calculate the expected value of exerting more effortful cognitive control (Shenhav et al., 2013). The ACC relies on input from multiple areas such as, sensory areas, the anterior insula (AI), amygdala, ventral-tegmental area (VTA), striatum and the PFC, to allocate control resources. The output, the execution of the effortful cognitive control, may then be conducted by the lateral PFC (Cai et al., 2016; C. Wang et al., 2016). Downstream projections from the ACC to areas such as the subthalamic nucleus and the LC, may also engage different types of control by modulating decision threshold and neural gain (Aston-Jones & Cohen, 2005; Cavanagh et al., 2011, 2014; Eldar et al., 2013; Jepma & Nieuwenhuis, 2011; Keuken et al., 2015). Furthermore, Westbrook et al. (2020) found that individuals with low dopamine synthesis capacity in the caudate nucleus showed higher cognitive effort costs. However, administering methylphenidate (increasing levels of dopamine and norepinephrine), increased their willingness to do cognitive work, showing that dopamine may promote cognitive work. Analyzing eye gaze patterns and computational modeling with a drift-diffusion model (see 1.5 Sequential sampling models of decision-making) dopamine biases attention towards benefits and away from costs. Phasic dopamine binding at striatal D1 receptors may be associated with benefits, whereas striatal D2 receptors may be associated with costs, thus reflecting benefits versus the cost of actions (Westbrook et al., 2020, 2021). Tonic ventral striatal dopamine may favor disengagement due to higher opportunity costs (Niv et al., 2007; see Westbrook et al., 2021 for a more detailed overview). And dopamine in the PFC may function to maintain the stability of working memory representations, thus facilitating effortful cognition (Westbrook & Braver, 2016). This framework is also consistent with a body of work showing that incentives can increase cognitive performance on tasks requiring cognitive control (Botvinick & Braver, 2015; Padmala & Pessoa, 2011). However, the nature of the cost of cognitive effort remains uncertain (Musslick et al., 2018).

Proposed explanations for cognitive effort costs include metabolic by-product accumulation or resource limitations, computational or representational capacity limitations, structural limitations and opportunity costs. On a biological level a proposed metabolic account of cognitive effort costs is that long term cognitive effort exertion may cause build-up of potentially toxic substances that need to be recycled (Holroyd, 2016). Indeed, Wiehler et al. (2022) found elevated levels of glutamate in the lateral PFC following daylong cognitive effort exertion. Indicating that rising cognitive effort costs could be attributed to a need to maintain glutamate levels within certain boundaries. Further, elevated cognitive effort costs may limit effort recruitment and bias decisions away from cognitively effortful tasks (Wiehler et al., 2022). An alternative proposal is that engaging in cognitively demanding tasks depletes energy resources such as blood glucose (Gailliot et al., 2007; Gailliot & Baumeister, 2007), however evidence suggests that this is not so (G. R. J. Hockey, 2011; Kurzban et al., 2013; Molden et al., 2012). The effects of glucose on effort and performance are better explained in terms of motivation (Inzlicht & Schmeichel, 2012). From an information-processing perspective, cognitive effort costs arise due to limits on shared representational and computational resources of the processing system. In this perspective cognitive control restricts interference due to crosstalk when different tasks compete to use the same set of representations (Feng et al., 2014; Garner & Dux, 2015; Musslick et al., 2016). The advantage of shared representations lies in the ability to generalize, draw inferences, and learn abstract structures (LeCun et al., 2015; McClelland & Rumelhart, 1985; Musslick et al., 2016). Thus, the advantage of shared representations outweighs the constraints on information processing (Feng et al., 2014; Musslick et al., 2016). Additionally, there are explanations focusing on structural capacity limitation which refer to restrictions in the number of computations that can be processed in a central control mechanism, often referring to limits of working memory or attention (Cowan et al., 2012; Kahneman, 1973; G. A. Miller, 1956). However, there is a lack of satisfactory explanations for why a structural limitation should have evolved. Lastly, it has been proposed that cognitive effort costs and restrictions on cognitive control arise due to an opportunity cost (Aston-Jones & Cohen, 2005; Kurzban et al., 2013). This account proposes that cognitive effort and the aversive feeling arises as the brain continually evaluates alternative uses of cognitive resources and alternative courses of actions. As one commits cognitive resources, time and action on a certain task, it excludes the engagement in alternative tasks. It has been proposed that the LC-NE system regulates the adaptive dilemma of exploiting and exploring the environment (Aston-Jones & Cohen, 2005), and that tonic striatal dopamine may signal the average reward rate or opportunity cost (Niv et al., 2007). Indeed, due to opportunity costs limiting computational effort and time on task can be the optimal solution (Gershman et al., 2015; Gigerenzer, 2008; Simon, 1990). Note that the accounts mentioned are not mutually exclusive.

The study of cognitive effort through decision-making paradigms such as the DST and COG-ED has significantly advanced cognitive effort research. However, until now there have been no studies evaluating whether these task paradigms are measuring the same cognitive effort construct, or if they are related. It has been shown that the effort discounting in COG-ED is related to NFC (Westbrook et al., 2013). However, demand avoidance in the DST is not related to NFC (Strobel et al., 2020). Further, the task paradigms have been applied to clinical research, showing disparate results regarding cognitive effort and schizophrenia (Culbreth et al., 2016; J. M. Gold et al., 2015). Thus, there is a need to compare tools measuring cognitive effort in order to determine their shared variance, i.e., to avoid misconceptions due to assumed similarities. Additionally, an alternative approach, not yet properly discussed, measures cognitive effort through items from the heuristics and bias literature. Several authors have noted that humans rely on heuristics, or simplifying rules, to save effort when making decisions (J. St. B. T. Evans, 2008; Frederick, 2005; Gigerenzer & Gaissmaier, 2011; Kahneman, 2011; Shah & Oppenheimer, 2008; Stanovich, 2016; Toplak et al., 2014; Trippas et al., 2015). It has long been assumed that many errors in reasoning and decision-making occur due to heuristics or fast effortless processing, and that engagement of more effortful cognitive processing leads to fewer errors in reasoning (J. St. B. T. Evans, 2008; Frederick, 2005; Kahneman, 2011; Stanovich, 2016; West et al., 2008). The assumption of effortless reasoning leading to errors in decision-making on bias and heuristics tasks have been so strong that performance on these tasks have been used as measures of deliberate (control demanding) processing and an analytic thinking style (Frederick, 2005; Pennycook et al., 2015a; Stanovich, 2016; Thomson & Oppenheimer, 2022; Toplak et al., 2014; Trippas et al., 2015). Supporting the notion that these tasks are related to effort and are measuring deliberate processing is the relationship between performance on these tasks and NFC (Thomson & Oppenheimer, 2016; Toplak et al., 2014; West et al., 2008). However, it is not known if performance on these tasks is related to other measures of cognitive effort, such as demand avoidance in the DST and effort discounting in COG-ED, which would be predicted by dual-process theories. As reviewed in Thomson & Oppenheimer (2022) there is a need to investigate cognitive effort across disciplines to advance knowledge across different fields of research. A possible approach to advance cognitive effort research is therefor by comparing tools and measures from different research fields.

1.2 Dual-process models of decision-making

Dual-process models have a long history in psychology and became more prominent in research on errors in reasoning and decision-making in the 1970's. Dual-process models have since been applied to other areas of research, such as social psychology (Strack & Deutsch, 2004), behavioral economics (Kahneman, 2011; Thaler & Sunstein, 2008), and moral philosophy (Białek & De Neys, 2017; Cushman, 2013; Greene, 2014). Dual-process models generally propose that human reasoning relies on two different modes of processing. Type 1 processing, often called intuitive or heuristic, is automatic and does not require working memory capacity. Type 1 reasoning is often associated with features such as being fast, effortless, unconscious, contextualized, associative and parallel (J. St. B. T. Evans & Stanovich, 2013; Kahneman, 2011; Stanovich & West, 2000). Type 2 processing, often called analytic or deliberate, relies on working memory resources to generate responses (J. St. B. T. Evans & Stanovich, 2013). Type 2 processing is often associated with being slower, effortful, conscious, decontextualized, rule-based, logical and serial (Epstein, 1994; J. St. B. T. Evans, 2006, 2008; Sloman, 1996; Stanovich, 2009a; Stanovich & West, 2000). At the center of the distinction between Type 1 and Type 2 processing is the difference in load on working memory. Type 2 processing taxes working memory resources, whereas Type 1 processing utilizes little to no working memory resources. However, there is disagreement on the nature of the two modes of processing, how they interact, and what responses they produce (De Neys, 2018; J. St. B. T. Evans & Stanovich, 2013).

The Default-interventionist account (J. St. B. T. Evans, 2008; J. St. B. T. Evans & Stanovich, 2013; Kahneman, 2011) propose that Type 1 processes are the default, and most situations generate intuitive responses. Type 2 processes are engaged (intervene) at later stages of reasoning, or not at all. Type 2 processes are treated as more computationally expensive (cost) but lead to the correct response more often (benefit). A trade-off between accuracy and computational expense is therefore evident (Simon, 1990; Stanovich, 2018). According to the Default-interventionist account, rational reasoning and sound decision-making is thus dependent upon Type 2 processing overriding errors made by Type 1 processing. In this account, humans are cognitive misers because their default is to conserve effort expenditure by relying on Type 1 processing. The Default-interventionist account and the distinction between Type 1 and Type 2 processing can be exemplified by a classical problem from the heuristics and bias literature showing base-rate neglect (De Neys & Glumicic, 2008; Kahneman & Tversky, 1973).

In this problem, participants are asked to decide which out of two groups a person picked at random most likely belongs to. Participants are provided two pieces of information.

1) They are given a description of the person. Example from Kahneman & Tversky (1973).

"Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to feel little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense."

2) They are provided information about how many people each group consists of (base-rate information). For this example, participants are told there are 900 lawyers and 100 engineers.

According to the Default-interventionist account, Type 1 processing will associate the description of Tom W. with a stereotypical engineer. According to Kahneman & Tversky (1973), the degree of representativeness (match between description and stereotype) will then be used to judge the probability that Tom W. is an engineer, as a Type 1 process. Thus, participants relying on Type 1 processing will neglect the base-rate information and respond that Tom W. is most likely an engineer. Integrating the prior probability given by the base-rates of engineers and lawyers (100 vs. 900), with the individuating information is thought to be a Type 2 process. Normatively, this results in Tom W. being more likely a lawyer due to the base-rates heavily favoring any person drawn from the two groups being a lawyer. Accordingly, Type 1 and Type 2 processing will have conflicting outputs. The Default-interventionist account proposes that most people will intuitively think that the description matches the stereotype of an engineer, however they would have to engage in deliberate effortful Type 2 processing to override the intuitive response and come up with the normative response by integrating both the base-rate and individuating information.

Parallel processing accounts propose that both Type 1 and Type 2 processing are engaged from the start of reasoning (Epstein, 1994; Sloman, 1996). Sloman (1996)'s two-systems theory proposes that there is one rule-based system (Type 2 processing) and one associative system (Type 1 processing). The rule-based system operates on symbolic structures and abstract variables. These symbolic structures have logical content and can assume a class of possible values. Rules are productive in that they can encode any number of propositions. Additionally, rules are systematic in that if they can be applied to one case, they can make inferences and apply to alternative cases. The associative system encodes and processes statistical regularities of its environment and computes on the basis of similarity and temporal contiguity. Both systems operate in parallel and can create different solutions to the same reasoning problem. The rule-based system can suppress the associative system but not completely inhibit it. Epstein (1994)'s cognitive-experiential personality theory proposes an experiential system (Type 1) and a rational system (Type 2). The experiential system is automatic and operates on classical learning principles, classical and operant conditioning, and observation. The experiential system solves problems automatically by reacting in accordance with reinforcement history. The rational system is a conscious reasoning system that evaluates evidence and applies logical principles when solving problems. Thus, for the parallel processing models it would be assumed that associating a description of a person with a stereotypical profession would be a Type 1 process (associative- or experiential system), whereas applying Bayes rule or otherwise applying rules to integrate the base-rate information with the individuating information would be a Type 2 process (rule-based- or rational system).

Both default-interventionist and parallel processing accounts assume that a Type 2 correction of a competing intuitive Type 1 response is necessary. Historically, these models have assumed that Type 2 processes are logical, rule-based, mathematical, or probabilistic, whereas Type 1 processes are heuristic, associative, stimulus-response pairings, and based on prior beliefs (not an exhaustive list) (Epstein, 1994; J. St. B. T. Evans, 2006, 2008; Kahneman, 2011; Sloman, 1996; Stanovich, 2009a; Stanovich & West, 2000). Often normative responses have been taken as evidence of Type 2 processing and errors on reasoning tasks have been assumed to result from Type 1 processing. Additionally, fast responses have often been assumed to result from Type 1 processing and slower responses from Type 2 processing (Stanovich & Toplak, 2012). However, evidence suggests that probabilistic and "normative" responses are often given fast and "intuitively" (Bago & De Neys, 2017; Newman et al., 2017; Raoelison et al., 2020). Furthermore, there is evidence that conflicting responses have been detected even when participants respond with the "intuitive" response. In response, there have been theoretical developments of dual-process models and a new generation of "hybrid" models or "dual-process 2.0" have been proposed (De Neys, 2018; De Neys & Pennycook, 2019; Pennycook et al., 2015b; Raoelison et al., 2020; Stanovich, 2009a, 2018; Thompson et al., 2018).

Stanovich (2009) proposes that the generic dual-process model should be refined with a three-process model where Type 2 processing needs to be divided into a reflective level and an analytical level. In Stanovich's (2009) three-process model the analytical level is responsible for carrying out mental simulation (or cognitive decoupling) in working memory. Meaning the generation of Type 2 responses and overriding of Type 1 responses are carried out at the analytical level. At this level individual differences in efficiency of processing are evident. The capacity of the analytical level can be assessed with traditional measures of intelligence. However, just as Type 1 processes need to be overridden by the analytical level (Type 2 process), the execution of this overriding is initiated at a higher level, namely the reflective level (Stanovich, 2009a). The reflective level regulates behavior at a high level of generality and is concerned with pragmatic- and epistemic self-regulation as well as higher order goals and values. Additionally, conflict detection between multiple intuitive responses, error monitoring, and the initiation of analytical (Type 2) processing are performed at the reflective level. According to Stanovich (2009), individual differences at the reflective level can be assessed with measures of thinking dispositions such as actively open-minded thinking (Stanovich & West, 1997) and NFC (Cacioppo & Petty, 1982), or with performance measures of rational reasoning such as the cognitive reflection test and other heuristics and bias tasks (Frederick, 2005; Stanovich, 2009a, 2016; Toplak et al., 2014; West et al., 2008). Importantly, individual differences in rational reasoning performance are thus dependent on both cognitive ability (intelligence) and thinking disposition (cognitive motivation).

Stanovich (2018) highlights that mindware instantiation, the degree to which mindware has been learned and automatized, is critical for understanding and classifying responses on rational reasoning tasks. The degree to which mindware is learned and automatized predicts whether conflict detection is possible, probable, or if a mindware specific response can be made intuitively as a Type 1 process[1] (Stanovich, 2018). Importantly, according to this framework a number of different intuitions can be produced as Type 1 responses. This depends on the degree of mindware instantiation, or the degree to which the specific mindware has been learned, practiced, and automatized. Importantly, Type 2 processes will not always lead to more accurate answers. If mindware is missing, deliberation will not lead to correct responding. Additionally, rationalization of an intuitive incorrect response can lead to a deliberate incorrect response (J. St. B. T. Evans, 2019; Pennycook et al., 2015b). The view of a multitude of intuitions being possible at the level of Type 1 processing is in line with a body of work which has found evidence for fast intuitive correct responses (Bago & De Neys, 2017; De Neys, 2018; Kruglanski & Gigerenzer, 2011; Mækelæ & Pfuhl, 2019; Newman et al., 2017; Raoelison et al., 2020; Thompson et al., 2018). This has given rise to the Smart intuitor account (Raoelison et al., 2020; Thompson et al., 2018).

The Smart intuitor account proposes that high cognitive capacity individuals are more likely to answer correctly on reasoning tasks by having “better” or more accurate intuitions (Raoelison et al., 2020). That means that a corrective deliberate process (as proposed by the Default-interventionist account) can still happen, but most of the variance in correct responding in decision-making tasks are explained by more accurate intuitions rather than overriding of faulty intuitions (Raoelison et al., 2020). Therefore, according to the smart intuitor account, overriding of incorrect intuitive responses is not always necessary on heuristic and bias tasks. Rather, most normative responses can be made intuitively. Thus, the smart intuitor account predicts that correct responses can be made fast and with little effort. Additionally, the smart intuitor account predicts that cognitive ability should predict reasoning performance, and cognitive motivation (NFC) should have less influence on performance. This is in contrast to the Default-interventionist account where both ability and motivation to execute deliberation is necessary for correct responding.

Addressing the question of how Type 2 processing is engaged is at the center of Pennycook et al. (2015b)’s three-stage model of analytic engagement. This model integrates the smart intuitor proposal

[1] Stanovich (2009) uses the term autonomous set of systems instead of Type 1 processing. This to highlight that there are multiple sub-systems working in parallel, not one unitary system.

that many types of intuitions can be made fast as a Type 1 process. In this model competing intuitions are the mechanism causing deliberation or Type 2 processing. The model suggests that bias can occur early as a failure in conflict monitoring, which leads to the initial response being given (similar to Default-interventionist). If a conflict between competing intuitions is detected Type 2 processing will be engaged. This can lead to evaluation of the initial responses, where the best one is given, or a new response can be generated. This is labeled cognitive decoupling, a type 2 process, similar to the proposal of the Default-interventionist account. Alternatively, if a conflict is detected, an initial incorrect response may be rationalized (Type 2 process) as being the correct one, that is similar to motivated reasoning (Kahan, 2013, 2015; Kunda, 1990).

Dual-process theories and heuristics and bias tasks have co-evolved where errors in reasoning have been explained in a dual-process framework and performance measures on heuristics and bias tasks such as response accuracy and response times have been used to investigate dual-process theories (De Neys & Glumicic, 2008; Kahneman, 2011; Kahneman & Frederick, 2002; Pennycook et al., 2015b, 2016; Pennycook, Fugelsang, et al., 2012; Stanovich, 2016; Stanovich & West, 2000; Tversky & Kahneman, 1974). A salient example is the base-rate task mentioned previously. Kahneman & Tversky (1973) presented evidence that humans tend to ignore base-rates in certain scenarios, although they are fully able to use base-rate information in the absence of individuating information. This error has since been explained as resulting from Type 1 processing (Kahneman & Frederick, 2002). However, evidence suggests that base-rate information can be used intuitively, and base-rate use depends on the task structure and the format of the information presented (Barbey & Sloman, 2007; Bar-Hillel, 1980; De Neys & Glumicic, 2008; Gigerenzer et al., 1988; Koehler, 1996; Pennycook & Thompson, 2012). Further the cognitive reflection test (CRT) has been proposed to assess cognitive reflection or the tendency to engage in Type 1 or Type 2 processing through task performance (Frederick, 2005). However, testing the famous bat and ball problem from the CRT in a two-response format, it was shown that most of the participants who answered correctly did so intuitively, and there was only a marginal increase in correct responding after deliberation (Raoelison et al., 2020). Further, a meta-analysis revealed that the CRT may assess general intelligence and numeracy, but no specific factor of cognitive reflection could be distinguished (Otero et al., 2022). Thus, there has been an assumption that task performance could display effort engagement and Type 2 processing without measuring actual effort in reasoning. Furthermore, errors in rational reasoning have been interpreted in dual-process frameworks, and proposed to occur due to fast effortless processing. A salient example of this is the teleological reasoning bias (Kelemen et al., 2013). Teleological reasoning is the tendency to see purpose and intentionality in natural phenomena. Teleological reasoning arises early in children's development and is applied as a general explanatory default (DiYanni & Kelemen, 2005). Later in life, through education a mechanistic, scientifically more accurate, explanation of natural phenomena replaces teleological reasoning. It has been assumed, in accordance with Default-interventionist dual-process theory that this developmentally persistent reasoning bias occurs as Type 1 processing, and has to be suppressed through Type 2 processing, in order to come up with an alternative more scientifically accurate explanation (Kelemen et al., 2013). This was supported by more teleological reasoning errors seen in a speeded vs. an un-speeded teleological reasoning task. Thus, it was assumed that Type 2 processing leads to less teleological reasoning errors. However, errors in the control condition were percentage wise equally affected by time pressure. This leaves the question of whether reasoning performance on these tasks are actually measuring Type 1 and Type 2 processing unanswered. Further, it is unclear if more effort and deliberation actually leads to better performance on these tasks. It is therefore important for both the development of dual-process theories and the understanding of errors in reasoning to distinguish the contributing factors to reasoning performance and Type 1 and Type 2 processing.

As the defining distinction between Type 1 and Type 2 processing is working memory load (J. St. B. T. Evans & Stanovich, 2013), using pupillometry (as the pupil is known to reflect cognitive effort and working memory load) is an excellent (and relatively cheap) way to investigate dual-process reasoning during task performance. Surprisingly, to our knowledge there is only one study that has investigated classic heuristics and bias tasks using pupillometry. Eldar et al. (2021) showed that larger pupil dilations were associated with more bias (not less as proposed by the Default-interventionist account) on three framing tasks and found no association between pupil dilation and bias on three other tasks. Therefore, there is a need to measure concurrent effort exertion during performance on a range of heuristics and bias task. Additionally, with new theoretical developments in dual-process research, it is important to make strong tests of the new frameworks in order to advance dual-process research.

1.3 Pupillometry as a measure of cognitive effort

The human pupil has a diameter that varies roughly between 2- and 8-mm (Mathot, 2018; Sirois & Brisson, 2014). The pupil constricts and dilates in response to brightness (controlling how much light enters the eye), fixation (controlling focus and visual acuity, constricting when changing from far to near and dilating from near to far), and mental activity (cognitive effort and arousal). Changes in brightness (from average lighting conditions to darkness) can more than double the pupils' usual size (approximately 120%), whereas changes due to mental activity are smaller and rarely more than 0.5 mm (Beatty & Lucero-Wagoner, 2000). Pupil dilation, in the context of a cognitive task, is usually measured as a stimulus induced change in pupil size from a pre-stimulus baseline time period. The pre-stimulus baseline pupil size is usually corrected for by subtraction or a divisive baseline correction (Mathot, 2018).

Starting in the 1960's a series of studies revealed that the pupil dilates in response to increasing processing demands across a range of different tasks such as, mental arithmetic, pitch discrimination, language comprehension and more (Boersma et al., 1970; Bradshaw, 1968; Hess et al., 1965; Hess & Polt, 1964; Kahneman & Beatty, 1966; Kahneman & Wright, 1971; Schaefer et al., 1968). The claim that the pupil reflects changes in cognitive effort was first made by Hess & Polt (1964). Since then, several authors and reviews have noted that pupil dilations during cognitive task performance can be used as a measure of cognitive effort (Beatty & Lucero-Wagoner, 2000; Just et al., 2003; Kahneman, 1973; Laeng & Alnaes, 2019; Mathot, 2018; van der Wel & van Steenbergen, 2018). The close relationship between cognitive effort and attention was early noted, Kahneman (1973) even used the terms interchangeably in his seminal book "Attention and effort", where "attentional effort" is used to describe the "intensity" aspect of attention, as opposed to the "selective" aspect of attention. Furthermore, attention and cognitive effort are closely related to the arousal system, which is covered in the next section (1.4 Pupillometry as a measure of the Locus Coeruleus – Norepinephrine system). Kahneman (1973) highlights that pupillometry is an adequate physiological measure of effort. Pupillometry is sensitive to changes in effort, between tasks, within tasks, as well as being sensitive to individual differences in effort requirements due to differences in processing capacity (cognitive ability). Kahneman (1973) noted that arousal and effort are not determined prior to the task but varies moment to moment when a subject is engaged in a task, and these variations correspond to changes in task demands. A salient example displaying changes in pupil size due to variations in cognitive effort can be seen in a digit-span task. In this task, subjects are presented a string of numbers to hold in memory and report back. The presentation of each successive digit in this task is followed by an increase in pupil dilation, with pupil size increases corresponding to increasing numbers of digits retained in memory (Kahneman & Beatty, 1966; Kreis et al., 2020). Further, as subjects report back the numbers retained in memory a decrease in pupil size can be observed, as an "unloading" of digits retained in memory unfolds (D. A. Johnson, 1971). Importantly, pupil dilation increases with higher cognitive load until cognitive capacity reaches maximum capacity, but excessive load or "overload" (9 – 13 digits or 125%) leads to a constriction of the pupil (Granholm et al., 1996; Kreis et al., 2020; Poock, 1973), likely indicating a withdrawal of effort, expected when extremely high task demands lead to disengagement (Brehm & Self, 1989). This provides convincing evidence that pupil dilation reflects cognitive effort rather than simply reflecting task demand. Furthermore, changes in pupil dilation have been used to reveal differences in reactive and proactive control strategies in a continuous performance task, which would not be possible if pupil dilation simply reflected task demand (Chatham et al., 2009). Lastly, performance is a function of both effort and ability. Accordingly, differences in cognitive ability have been shown to influence pupil dilation during task performance (Ahern & Beatty, 1979; Bornemann et al., 2010; van der Meer et al., 2010). The evidence suggests that for simpler or routine tasks, more intelligent individuals have smaller pupil dilations during task performance, presumably due to higher efficiency of processing and thus lower effort expenditure needed to complete the tasks (Ahern & Beatty, 1979). Conversely, on difficult tasks it appears that individuals of higher intelligence have larger pupil dilations during task performance, presumably due to more available cognitive resources (Bornemann et al., 2010; van der Meer et al., 2010). However, it should be noted that the research literature regarding intelligence influencing task-evoked pupil responses is scarce, and these findings should be interpreted as preliminary. The relationship between cognitive effort and pupil size though, is well documented on a broad range of tasks, including but not limited to; memory tasks involving digit strings and memory recall, cognitive control tasks such as continuous performance and n-back, inhibition tasks such as go-/no-go, Stroop, and other conflict paradigms, mathematical problems, listening effort tasks manipulating complexity and intelligibility, sentence- and language comprehension, and mental spatial rotation (Beatty & Lucero-Wagoner, 2000; Just et al., 2003; Kahneman, 1973; S. E. Kramer et al., 2013; Laeng & Alnaes, 2019; Mathot, 2018; Piquado et al., 2010; van der Wel & van Steenbergen, 2018; Zekveld et al., 2018). However, the pupil does not only reflect changes in cognitive effort expenditure. As noted by Laeng et al. (2012), pupillometry may provide insight into changes in mental states, allocation of attention and the intensity aspect of mental activity, and thus could be considered a window to the preconscious. Of particular importance is the close relationship between pupil diameter and activity in the Locus Coeruleus, a brain stem nucleus involved in arousal, sleep, stress, attention, learning, behavioral flexibility, decision-making and more (Arnsten, 2000; Aston-Jones et al., 1999, 2001; Aston-Jones & Cohen, 2005; Berridge & Waterhouse, 2003; Bouret & Sara, 2005; Foote et al., 1975; Nassar et al., 2012; Poe et al., 2020; Preuschoff et al., 2011; Rajkowski et al., 1994; Sara & Segal, 1991; Yu & Dayan, 2005).

1.4 The Locus Coeruleus – Norepinephrine system, neural gain and pupillometry

The Locus Coeruleus (LC) is a small nucleus deep in the brainstem that synthesizes the neurotransmitter norepinephrine (NE)[2]. The LC has widespread effects on the central nervous system and the whole organism through projections to the spinal cord, brainstem, cerebellum and the forebrain (Aston-Jones et al., 1984; Berridge & Waterhouse, 2003; Foote et al., 1983; Moore & Bloom, 1979; Poe et al., 2020; Samuels & Szabadi, 2008). The primary source of cortical NE is the LC (Foote et al., 1983). The LC-NE system has far-ranging connections reaching most of the cortex through axonal varicosities with only few areas that do not receive innervation (Agster et al., 2013; Jones et al., 1977; Jones & Yang, 1985; Poe et al., 2020; Samuels & Szabadi, 2008). The LC appear to function both as a global unified brain-wide signal and modularly with functionally discrete actions (Chandler et al., 2014; Foote et al., 1983; Nagai et al., 1981; Schwarz et al., 2015; Uematsu et al., 2017). It has long been known that NE modulates global brain states such as arousal, alertness and the sleep-wake cycle (Arnsten, 2000; Aston-Jones et al., 1999; Berridge & Waterhouse, 2003; Foote et al., 1983; Poe et al., 2020). However, recent work suggests that clusters of LC neurons seem to be organized with projections to functionally related target areas such as the amygdala and mPFC, coordinating for example aversive learning and behavioral flexibility (Chandler et al., 2014; Uematsu et al., 2017). Additionally, the LC-NE system is involved in sensory processing, memory formation, gene-transcription and brain plasticity, executive functions, such as working memory, focused and flexible-attention, in addition to cognitive- and behavioral flexibility and decision-making (Berridge & Waterhouse, 2003; Bouret & Sara, 2004; Chandler et al., 2014; McBurney-Lin et al., 2019; McGaughy et al., 2008; Poe et al., 2020; Sara & Bouret, 2012; Spencer & Berridge, 2019; Takeuchi et al., 2016; Viglione et al., 2023; Waterhouse & Navarra, 2019). Accordingly, there are multiple factors affecting the influence of NE which regulates cognition and behavior differentially at different levels and timescales.

The LC-NE system primarily functions as a neuromodulatory system, that is, NE acts through modulating the effects of other neurotransmitters such as glutamate and gamma amino butyric acid (GABA), rather than producing direct excitatory or inhibitory effects. Thus, the LC-NE system modulates the strength and efficiency of synaptic transmission and the overall activity pattern of neural circuits. It has been proposed that the LC-NE system globally functions through modulating neural gain (Aston-Jones & Cohen, 2005; Eldar et al., 2013; Servan-Schreiber et al., 1990; Usher et al., 1999). That is, the LC-NE system alters the responsiveness of neural populations to synaptic input,

[2] Although all Locus Coeruleus neurons contain norepinephrine some also express other molecules.

either increasing or decreasing the signal based on other input. Higher gain increases the signal-to-noise ratio in neural circuits by amplifying salient information or representations, while suppressing other information. Global increases in neural gain are associated with focused attention, higher functional connectivity, and clustering of neural activity, whereas low gain is associated with broader attention and lower functional connectivity (Eldar et al., 2013). According to the adaptive gain theory (Aston-Jones & Cohen, 2005) the LC- NE system adaptively adjust gain to the adaptive dilemma of exploiting and exploring the environment.

Importantly, LC activity can be characterized as tonic and phasic. Tonic LC activity is the baseline firing rate of the LC and is monotonically related to wakefulness and arousal. Phasic LC activity is characterized by short bursts of spiking neural activity. It occurs at moderate levels of tonic activity and is often linked to salient or novel stimuli. According to the adaptive gain theory, phasic LC activity creates system-wide increases in neural gain to facilitate the execution of behavioral responses ensuing from decision processes. In the context of experimental tasks, phasic activity is associated with task related stimuli but not distractors. Low tonic LC activity i.e., low NE levels, are associated with inattention, drowsiness, little motor activity and low cognitive performance. Moderate tonic activity is associated with phasic activity, high cognitive performance, attentiveness and wakefulness. High tonic LC activity (or tonic mode) has been associated with distractibility and low cognitive on-task performance. According to the adaptive gain theory, tonic mode produces an enduring non-specific increase in gain that increases task disengagement, thus task performance suffers. However, the exploration of alternatives is an adaptive adjustment on a broader scale and necessary when utility in the current task decreases. The relationship between NE and task performance can be described as an inverse U-shaped function. An explanation for this can be found when considering that NE acts at three families of receptors, α1, α2, and β receptors. α2 receptors have higher affinity to NE compared to α1 and β receptors. At low and moderate levels of NE, high affinity α2 receptors promote working memory. However, at higher levels of NE (e.g., high arousal conditions such as stress and anxiety) lower affinity α1 receptors can impair working memory performance (Arnsten, 2000; Berridge & Spencer, 2016; MacDonald et al., 1997; M. Wang et al., 2007). However, low affinity α1 receptors promote both focused- and flexible attention at moderate and higher levels of NE, but also show diminished performance with levels of NE being too high. Thus, differing levels of NE enhance and impede separate executive cognitive processes. Notably sustained and flexible attention is enhanced at higher levels of NE compared to working memory, thus indicating a right-shifted inverted U-curve (Berridge & Spencer, 2016; Spencer & Berridge, 2019).

Alternative theories have proposed that LC activity may signal unexpected uncertainty or environmental volatility, and learning dynamics of the environment (Browning et al., 2015; Dayan & Yu, 2006; Nassar et al., 2012; Preuschoff et al., 2011; Yu & Dayan, 2005). Phasic LC activation may function as a neural interrupt signal or a “network reset”. Phasic LC activation occurs in response to a change in the environment (salient, novel, or unexpected stimulus or event), and the function of this signal might be to interrupt ongoing activity and facilitate re-organization, in order for fast behavioral adaptation to occur (Bouret & Sara, 2005; Dayan & Yu, 2006).

There is a remarkably high correlation between LC activity and fluctuations in pupil diameter (Joshi et al., 2016; Rajkowski et al., 1994; Reimer et al., 2016). Thus, non-luminance mediated changes in pupil size as measured by pupillometry can be used as a proxy for the LC-NE system and neural gain (Eldar, Cohen, et al., 2016; Eldar et al., 2013; Eldar, Niv, et al., 2016; Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; Joshi et al., 2016; Murphy et al., 2014). Notably, there is an inverse relationship between baseline pupil size and task evoked pupil responses (de Gee et al., 2014; Eldar et al., 2013, 2021; Gilzenrat et al., 2010; Murphy et al., 2014). Meaning lower baseline pupil size, indicating low levels of NE and neural gain, can be associated with larger pupil dilations. Thus, indicating a need for caution when making baseline corrections. Furthermore, using pupil dilation as a proxy for neural gain, and modelling the influence of neural gain on decisions, has received attention in an alternative framework for decision-making, namely sequential sampling models (Busemeyer et al., 2006; Eldar et al., 2013, 2021; Krajbich & Rangel, 2011; Usher et al., 1999, 2013; Usher & McClelland, 2004).

1.5 Sequential sampling models of decision-making

Sequential sampling models of decision-making are a class of cognitive computational models that aims to understand and describe the cognitive processes underlying decision-making (Boag et al., 2023; N. J. Evans & Wagenmakers, 2020; B. U. Forstmann et al., 2016; Ratcliff et al., 2016). Generally, these models assume that decision-making is a process which unfolds over time as evidence is gradually accumulated and integrated, at a certain rate, for some alternatives, until a threshold or criteria is reached, which triggers a decision. These models are formally described mathematical models which provide predictions regarding both decision choice and response times for decisions. An advantage of these models is the ability to decompose response time distributions into latent psychological parameters underlying decisions, such as drift-rate and decision threshold, rather than relying on descriptive statistics such as mean response time and percentage of correct answers. Thus, the models can account for speed-accuracy trade-offs present in most decision-making tasks. Further, sequential sampling models enable researchers to link latent psychological decision-parameters with psychophysiological data (Cavanagh et al., 2014; de Gee et al., 2020; Krajbich, 2019; Krajbich et al., 2010, 2012; Turner et al., 2013, 2015, 2016; Westbrook et al., 2020). These models are neurally plausible both in individual neurons and populations of neurons (Arnold et al., 2015; Ding & Gold, 2012; B. U. Forstmann et al., 2016; J. I. Gold et al., 2008; J. I. Gold & Shadlen, 2007; Roxin & Ledberg, 2008; Shadlen & Kiani, 2013). Sequential sampling models can theoretically be extended to any cognitive decision-making task and have been applied and validated with a range of cognitive tasks in areas such as perceptual choice, learning, memory, language processing and consumer choice (B. U. Forstmann et al., 2016; Krajbich, 2019; Lerche et al., 2020; Lerche & Voss, 2017, 2019; Ratcliff, 1978; Ratcliff et al., 2016; Trueblood et al., 2014; Voss et al., 2004; Westbrook et al., 2020). To exemplify the general structure of sequential sampling models, the drift-diffusion model (DDM, also known as diffusion decision model) will be presented (Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998; Wagenmakers et al., 2008). The model has four key parameters, drift-rate, threshold, starting point, and non-decision time. The model assumes two choice options and that evidence is noisily and gradually accumulated from a starting point until a decision threshold is reached. The drift-rate represents the average amount of evidence for the two response options per unit of time. Stronger evidence in favor of one option will lead to higher drift-rate, whereas weaker or more ambiguous evidence will lead to lower drift-rate. Thus, drift-rate is indexing task difficulty. Additionally, individual ability, intelligence or processing speed will also influence the rate of evidence accumulation or drift-rate. Further, two decision thresholds, representing each response option, set the criteria for when the evidence is sufficiently favoring one option. Thus, evidence is accumulated until it favors one option to an extent that it reaches a decision threshold, triggering a decision. The decision-threshold thus indicates response caution (conservativism) or impulsivity, and accounts for speed-accuracy trade-offs. Larger decision boundary separation requires that more evidence must be collected (or favor more strongly) one response option. This results in fewer errors, but longer response times. Conversely, lower decision thresholds lead to faster decisions but more errors. The starting point of the model reflects prior bias or preference for one of the response options. The starting point of evidence accumulation can be located closer to one boundary, and thus further away from the other response alternative, requiring less evidence accumulation for one option. The fourth parameter is non-decision time, such as sensory processing of the stimuli and execution of motor responses. Additionally, the model can include variability in drift-rate, starting point, and non-decision time across trials (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002).

Several sequential sampling models have been proposed and they differ from the DDM in various ways. Such as, whether evidence is accumulated at discrete time points or continuously, if there is one or more accumulators, whether these are independent, if they are leaky, or exert influence on the accumulation process, whether thresholds are fixed for each accumulator or depend on the relative evidence strength, and if decision boundaries are static or collapsing over time (Bogacz et al., 2006; Brown & Heathcote, 2008; Ratcliff, 1978; Smith & Ratcliff, 2004; Smith & Vickers, 1988; Teodorescu & Usher, 2013; Usher & McClelland, 2001; Van Zandt et al., 2000; Vickers & Lee, 1998). Additionally, there are differing theoretical accounts of the decision-making process (Brown & Heathcote, 2008; Busemeyer & Townsend, 1993; Krajbich & Rangel, 2011; Ratcliff, 1978; Ratcliff & Rouder, 1998; Roe et al., 2001; Usher & McClelland, 2001). These sequential sampling models propose different dynamics of the decision-making process and have different applications. For example, the attentional drift-diffusion model proposes that attention influences the rate of evidence accumulation for the option being attended (Krajbich, 2019; Krajbich et al., 2010). The model explains how attention is dynamically allocated and how visual attention dynamically influences decision-making. The model can be extended to include multiple options and be applied to domains such as valuation, memory, and learning. Studies using eye-tracking reproduce predictions from the model, such as the chosen option being attended more, and usually last before making a decision. Relatedly, decision field theory proposes that individuals’ preferences change over time as different aspects of options are considered and evaluated, influencing the perceived value of the options and the preference state of the options (Busemeyer & Townsend, 1993; Diederich, 2003; J. G. Johnson & Busemeyer, 2005; Roe et al., 2001). Thus, highlighting the role of the dynamic process evolving over time, and the influence of context, which will influence attention, valence and preference-state. Additionally, the model can be extended to include outcomes that are valuations, further extending the possible applications of the model. Decision field theory and models built on this framework have been used to explain decision biases such as preference reversals and loss aversion (Busemeyer &

Townsend, 1993; Diederich, 2003; J. G. Johnson & Busemeyer, 2005; Roe et al., 2001). However, despite their differences many of these models have a high degree of overlap and they provide an accurate description of response time distributions in a number of cognitive tasks. Thus, it can be hard to distinguish predictions and assess which dynamics best explain the decision-making process (Dutilh et al., 2019; Leite & Ratcliff, 2010; Teodorescu & Usher, 2013).

Notably, it has been proposed that neural gain influences evidence accumulation in sequential sampling models (Aston-Jones & Cohen, 2005; Eldar, Cohen, et al., 2016; Eldar et al., 2013, 2021; Usher et al., 1999). Higher neural gain leads to higher weighting of each piece of evidence, such that fewer pieces of evidence are required to reach a decision-threshold and a decision. Conversely, low neural gain leads to lower weighting of each piece of evidence such that more evidence must be accumulated before a decision is made. Additionally, it has been proposed that neural gain influence the breadth of information processing such that high gain is associated with narrower attention, whereas low gain is associated with broader information sampling (Eldar et al., 2013; Eldar, Niv, et al., 2016). Thus, low gain allows a broader range of information to influence the evidence accumulation process. Importantly, Eldar et al. (Eldar et al., 2021) proposes that decision-biases occur due to more extensive integration. The authors highlight that predictions from more extensive integration due to low neural gain make the opposite predictions regarding pupil dilation and decision bias, as dual-process theories do. Indeed, their result supports that larger pupil dilations due to low neural gain, as proposed by extensive integration, is linked to more bias on a range of decision-making tasks (Eldar et al., 2021). As this study was the first to apply pupillometry to directly assess these competing decision-making frameworks, there is a need to further test this hypothesis with other reasoning bias tasks.

1.6 Aims of the thesis

As outlined above cognitive effort is ubiquitous, however the tools measuring cognitive effort have been developed in different strains of research and little is known about whether they are measuring the same construct. There is therefore a need to assess the shared variance of these tools to advance research and avoid errors due to mistaken similarity. Further, the role of cognitive effort in decision making and reasoning errors is controversial with recent empirical findings opposing predictions of established dual-process theories. Moreover, newly developed theoretical models are in need of empirical testing and alternative decision-making paradigms may explain current empirical findings. It is therefore pertinent to empirically test opposing decision-making frameworks and further shed light on the role of cognitive effort in decision-making and errors in reasoning.

The overarching aim of this thesis was to (1) investigate the role of cognitive effort in decision-making and errors in reasoning. To achieve this, I also (2) evaluated tools measuring cognitive effort in decision-making, and (3) evaluated competing dual-process models and alternative frameworks of decision-making.

The primary aim of Paper 1) was to evaluate tools measuring cognitive effort from different strains of research. A secondary aim of Paper 1 was to evaluate if other cognitive effort measures were related to performance on a rational reasoning tasks battery, as predicted by dual-process theory. We present six studies comparing different tools measuring cognitive effort to assess their shared variance. Thus, filling a critical gap in existing knowledge about the shared variance of cognitive effort measures within and across research fields. Additionally, we test an important assumption of dual-process theories, namely that cognitive effort expenditure is related to performance on rational reasoning tasks.

The aim of Paper 2) and Paper 3) was to investigate the role of cognitive effort in decision-making and test predictions of decision-making frameworks explaining bias in reasoning. To achieve this, I applied a psychophysiological measure of cognitive effort, pupillometry, during task performance on two different reasoning tasks. In Paper 2) I simultaneously recorded choices, response times and pupil dilation during performance on a teleological reasoning task. Thus, advancing our knowledge of a developmentally persistent reasoning bias, while explicitly testing competing decision-making frameworks and investigating the role of cognitive effort in reasoning bias. In Paper 3) I applied two variants of a well-established reasoning task, the base-rate task, adapted for eye-tracking and pupillometry, respectively. Measuring eye-tracking and pupillometry during task performance allowed me to investigate predictions from the three-stage model of analytic engagement (Pennycook et al., 2015b). Additionally, by applying psychophysiological measures and computational modelling of response times I bring insight into task performance on this reasoning task which have had significant influence on the development of dual-process models and our understanding of bias in reasoning.

2 Methods

Table 1 present an overview of the studies, samples, tasks, and measures applied in the three papers.

Table 1. Overview of studies, samples, tasks, and measures for the three papers in the thesis

Study	Sample	Tasks and measures	Pupillometry
Paper 1			
Study 1	WashU (N = 76)	DST, COG-ED, NFC, [a]	
Study 2	WashU (N = 91)	COG-ED, NCS, [a]	
Study 3: Day 1	UiT (N = 82)	COG-ED, N-TLX$_{\text{COG-ED}}$[b]	
Study 3: Day 2	UiT (N = 84)[c]	RQ, N-TLX $_{\text{RQ}}$, NCS	
Study 4: Day 1	UiT (N = 40)	DST, RQ, [a], NCS, [a], [a]	
Study 4: Day 2	UiT (N = 40)	DST, NCS, [a], COG-ED, N-TLX$_{\text{COG-ED,}}$	

Study 5: Day 1	UiT (N = 45)	DST, N-TLX$_{DST}$, RQ, N-TLX$_{RQ}$, NCS, Teleological reasoning task[d]	Teleological reasoning task[d]
Study 6: (1)	UiT (N = 91)	COG-ED, N-TLX$_{COG\text{-}ED,}$ DST, N-TLX$_{DST,}$ RQ, N-TLX$_{RQ}$, NFC	
Study 6: (2)	Prolific (N = 227)		
		Paper 2	
Study 5: Day 1	UiT (N = 45)	DST[e], N-TLX$_{DST}$[e], RQ, N-TLX$_{RQ}$, NCS, Teleological reasoning task	Teleological reasoning task
		Paper 3	
Study 5: Day 2[f]	UiT (N = 40)	Base-rate task (1)[g], N-TLX$_{BR(1)}$, RQ, Base-rate task (2), N-TLX$_{BR(2)}$	Base-rate task (1) Base-rate task (2)
+ Sample[h]	UiT (N = 20)	Base-rate task (1), N-TLX$_{BR(1)}$, RQ, Base-rate task (2), N-TLX$_{BR(2)}$, NFC	Base-rate task (1) Base-rate task (2)

Note. Data from Study 5: Day 1 was applied in both Paper 1 and Paper 2. Study 6 included two samples, the first sample (1) was collected at UiT and the second sample (2) was collected online through Prolific. Abbreviations: WashU = Washington University in St. Louis, UiT = UiT–The Arctic University of Norway, DST = Demand selection task, COG-ED = Cognitive effort discounting paradigm, RQ = Rational reasoning battery, NFC = Need for cognition scale, N-TLX = NASA task load index, BR = Base-rate task.

[a] Notes a task participants completed but not discussed in the thesis.

[b] NTL-X followed by task name in subscript = N-TLX for the task noted in subscript.

[c] In Study 3, N = 65 participants were tested on both days.

[d] Teleological reasoning task not reported in Paper 1.

[e] DST and N-TLX not reported in Paper 2.

[f] Participants from Study 5: Day 1 were invited for a second day of testing. Due to the COVID-19 pandemic not all participants could partake in Day 2.

[g] For the base-rate task (1) indicates the task version performed first, (2) indicates second.

[h] An additional sample was recruited for Paper 3 to achieve statistical power.

2.1 Studies, participants, and ethics

All three papers in this thesis were based on six studies. Two of the studies, Study 1 and Study 2 were conducted in 2013 at Washington University in St. Louis by an independent research group. The remaining studies, 3 – 6. were conducted at UiT – The Arctic University of Norway (UiT) between 2018 and 2022. Paper 1 was based on data from studies 1 – 6 (not included in Paper 1 is Study 5: Day

2, and the teleological reasoning task from Study 5: Day 1). Paper 2 was based on data from Study 5: Day 1 (not included in Paper 2 is the DST and N-TLX). Paper 3 is based on data from Study 5: Day 2.

For Study 1 and Study 2, all participants were tested individually at Washington University in St. Louis. Participants received 10$ per hour for participation, in addition to monetary rewards which could be gained in the COG-ED.

Study 3 consisted of two days of testing approximately three weeks apart. Participants were undergraduate students at UiT, receiving course credit for participation. Participation was possible for either one of the days, or both days. Testing was conducted in small groups in a computer room.

Study 4 consisted of two days of testing. Testing sessions were 4 – 8 weeks apart. All participants completed both days. Participants were a mix of students at UiT, full-time workers and high-school students. Participants received 200 NOK for participation and earned between 50 – 150 NOK (1 USD is approximately 10 NOK) for performance on COG-ED and a second task.

Study 5 consisted of two days of testing. For Paper 1, Study 5: Day 1 tasks: DST, N-TLX $_{\text{DST}}$, rational reasoning battery, N-TLX $_{\text{Rational reasoning battery}}$, and NCS, were used. For Paper 2, Study 5: Day 1 tasks: rational reasoning battery, NFC, and teleological reasoning task, were used. For paper 3, data from Study 5: Day 2 were used. Participants in Study 5 received either 400 NOK for two test days or 150 NOK for participating on only one of the days. Participants were a mix of students and full-time workers. Participants from Study 5: Day 1 were invited back for a second day of testing. However, due to the COVID-19 pandemic the testing session was delayed, and some participants could not partake in Study: 5 Day 2. To achieve the desired sample size, an additional sample was recruited to partake in Study 5: Day 2. This additional sample was further increased to account for participants from Study 5: Day 1 who could not partake in the second day of testing.

Study 6 consisted of two samples. The first sample (1) was undergraduate psychology students recruited from two-psychology classes at UiT. Participants received course credit for participation and could win vouchers worth approximately 25 – 50 NOK based on performance in COG-ED. The second sample (2) was collected online from Prolific (prolific.co). Participants at Prolific received 8 GBP for participation, plus bonus based on performance on COG-ED. Both samples completed the testing online.

Study 1 and Study 2 were approved by the Institutional Review Board at Washington University in St. Louis. All participants provided written consent. Study 3 - 6 were separately approved by the Institutional Review board at the Department of Psychology at UiT. All participants provided consent (written or online) before participation. All participants were 18 years or older, and younger than 65 years.

2.2 Tasks and measures

2.2.1 Cognitive effort discounting paradigm (COG-ED)

In paper 1, the cognitive effort discounting paradigm was used in all studies except for Study 5. The COG-ED measures differences in cognitive effort costs (Westbrook et al., 2013). The COG-ED relies on the n-back task (see Owen et al., 2005). In this task, participants are presented sequentially a series of letters in the middle of the screen. The task is to respond (key press) if the current letter is the same as the letter "N" numbers before. I.e., if N = 1 participants respond if the current letter is the same as the previous letter (current - 1). If N = 3 participants respond if the current letter is similar to the letter three letters back (current – 3). Participants are familiarized with the N-back task by playing all load levels for three runs. In Study 1 and Study 2 load levels were between N = 1 and N = 6. In Studies 3,4 and 6, the load levels were between N = 1 and N = 4. In the choice phase of the COG-ED participants are presented a series of decisions between performing an easy task (N = 1) for a smaller reward, or a harder task (N = 2 – 6) for a larger reward. see Figure 1 for an illustration of the N-back task and choice (adapted from Mækelæ et al., 2023).

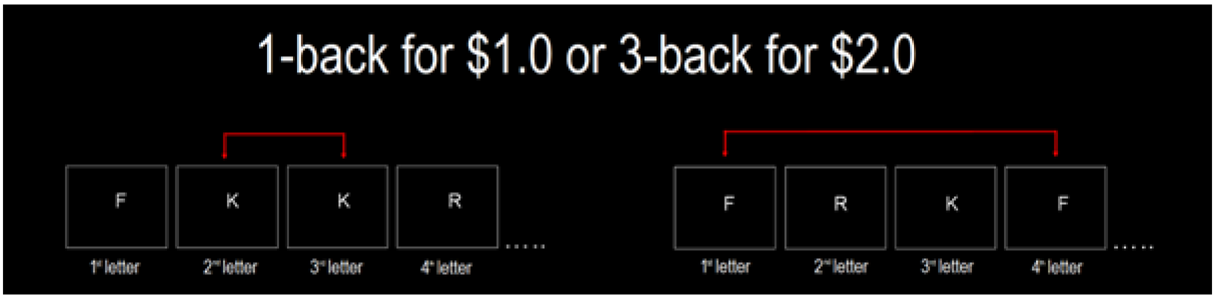


Figure 1. Illustration of the cognitive effort discounting paradigm.

Note. The figure illustrates the logic of the n-back task and the cognitive effort discounting paradigm. Example includes an illustration of a choice between performing a N = 1 back task for a 1$ reward or performing a N = 3 back task for a 2$ reward. Figure is adapted from Mækelæ et al. (2023).

Reward amount offered for the two options are adjusted and titrated based on participants choices to find the indifference point between two choice options. As an example, in Study 1 participants were offered 2$ for N = 2 and 1$ for N = 1. If participants chose the low effort option (N = 1), the reward amount for the low effort option was adjusted downward on the next choice. If the high effort option (N = 2) was chosen the reward amount for N = 1was adjusted upward on the next choice. After each choice the adjustment of the reward amount is halved. The final amount chosen after six choices is taken as the subjective indifference point between the two options. Subjective indifference points were averaged across all load levels to create an average indifference point score for each individual. This is the cognitive effort discounting measure for COG-ED in Paper 1.

2.2.2 Working memory capacity

Working memory capacity in Paper 1 was assessed by the discriminability score (d') of participants in the practice phase of n-back task in COG-ED. The calculation is based on signal detection theory.

Correct responses to a stimulus presented N-numbers back is considered a “hit”. Not responding to this stimulus is considered a “miss”. Not responding when the current letter does not match the letter N-numbers back is considered a “correct rejection”. “False alarms” are responses too late or too early. The discriminability score is calculated by z-transformed hit-rate minus z-transformed false alarm rate (d’ = z(H)-z(FA)). Higher discriminability score indicates higher working memory capacity.

2.2.3 Demand selection task (DST)

The demand selection task was used in Studies 1, 4, 5, and 6, in Paper 1. The DST measures cognitive demand avoidance. The task used was a replication of experiment 3 in Kool et al. (2010). In this task paradigm participants perform two different tasks. In both tasks participants are presented with a digit on the screen, digits range from 1 to 9, excluding 5. One task is a magnitude task where participants must respond to indicate if the digit is higher or lower than 5. The other task is a parity task where participants must indicate if the digit is odd or even. Which of the two tasks participants have to perform is indicated by the color of the digit, either blue or yellow. Responses are made by a mouse click on the right or left side of the screen to indicate choices. Participants start with a practice phase of 60 trials where participants are familiarized with the two tasks and the associated color. In the practice phase participants receive feedback and can redo the training phase if necessary (none of the participants had to redo the practice phase). In the test phase participants are presented with two colorful balls on the screen and should pick one of the two balls. See Figure 2 for an illustration of the demand selection task (adapted from Mækelæ et al., 2023).

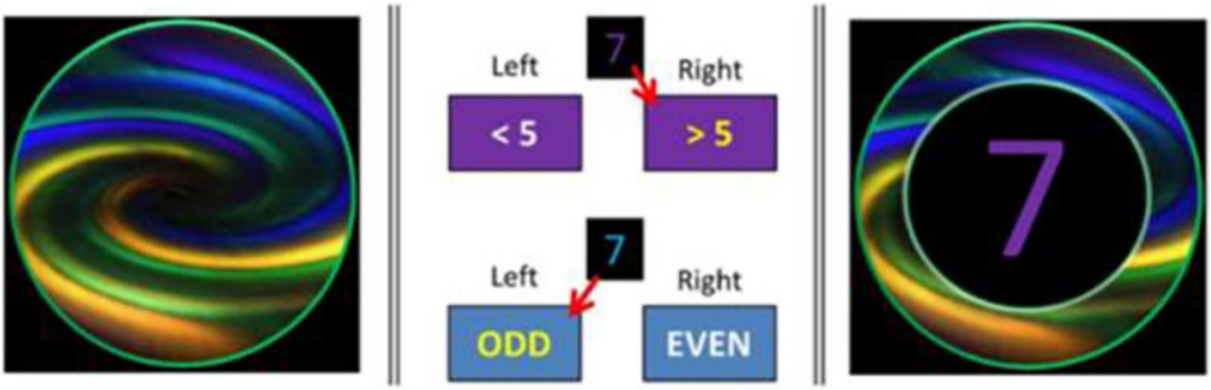


Figure 2. Illustration of the demand selection task.

Note. Figure illustrates stimuli and tasks. Left side of figure shows a ball (before choosing). Right side is the ball with number displayed (after choosing ball). Middle shows a magnitude task (top) and a parity task (bottom). Figure is adapted from Mækelæ et al. (2023).

Participants are instructed that they should sample from both presented options but can continue with one of them if they develop a preference. The location of the balls change (appearing along an invisible circle at a 45-degree angular distance). Not known by the participants one of the balls switches tasks on 90% of trials, and the other switches tasks on 10% of the trials. There are eight runs of 75 trials each, 600 trials total (Study 6 used 4 runs with a total of 300 trials). Cognitive demand avoidance is quantified as the proportion of high demand choices (choosing the ball with 90% chance of task switch). Thus, higher proportion of high demand choices (score between 0.5 and 1) indicates a preference for cognitive effort and conversely a lower proportion of high demand choices (score between 0 – 0.5) indicates cognitive demand avoidance.

2.2.4 Rational reasoning battery

The rational reasoning battery was used in studies 2 – 6 and both Paper 1 and Paper 2. The rational reasoning task battery consisted of a collection of tasks from the heuristics and biases literature with some variation in tasks across the studies. The tasks are considered a behavioral measure of a thinking disposition or a preference for cognitive reflection (or analytic cognitive style) and cognitive effort. Performance on the tasks is believed to reflect usual cognitive effort engagement, as opposed to maximum (Stanovich, 2009b). Higher performance indicates a tendency to engage in cognitive effort when necessary, and lower performance is associated with cognitive miserliness (Toplak et al., 2014; Trippas et al., 2015). Tasks from the cognitive reflection test (CRT) are applied in the task battery (Frederick, 2005; Thomson & Oppenheimer, 2016; Toplak et al., 2014). Performance on these tasks is believed to depend on the ability to suppress an intuitive but incorrect response in order to come up with a more deliberate correct response. The underlying assumption is that the deliberation process requires more cognitively demanding processing, thus correct responses reflect more cognitive effort (Frederick, 2005; but see Raoelison et al., 2020). As an example, consider problem 2 from the original CRT, "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?___ minutes." (Frederick, 2005, p. 27). Intuitively the answer 100 comes to mind for many people. Perhaps assuming that the number of machines, widgets and minutes should be the same, as they were so in the premise. However, if one takes the time to think through the problem, many people realize that it only takes each machine 5 minutes to create a widget, and the answer must be 5 minutes, as the number of machines and widgets are the same (however, see Raoelison et al., 2020).

Multiple tasks in the rational reasoning battery such as belief bias in syllogistic reasoning (Markovits & Nantel, 1989; Toplak et al., 2014) are also based on the premise that Type 2 processing must suppress responses created by Type 1 processing. Thus, these problems include an intuitive "lure". Additionally, there are tasks in the rational reasoning battery which have no "lure" option. For example the knight and knave problems (Smullyan, 1978) which depend on thinking through all options or fully disjunctive reasoning. These tasks are also believed to assess cognitive effort as they demand cognitive work find that there is a correct solution.

In addition to the tasks already mentioned we used a range of tasks including 18 items from Toplak et al. (2011), which includes the original CRT (Frederick, 2005). Further, items from the extended and alternative CRT were used (Thomson & Oppenheimer, 2016; Toplak et al., 2014), a "marriage problem" (Levesque, 1986), a conditional reasoning problem (Lehman et al., 1988), a medical statistical reasoning problem (Gigerenzer et al., 2007), and a base-rate problem (West et al., 2008).

Correct responses on each problem were scored as 1, and incorrect responses were scored as 0. In paper 1, the proportion of correct responses were calculated to compare scores across studies. In paper 2 a composite score between 0 – 14 was created for each individual and subsequently z-scored across participants.

2.2.5 Need for cognition scale (NFC)

Need for cognition was used in all studies, and included in both Paper 1 and Paper 2. NFC was measured with the abbreviated 18-item version (Cacioppo et al., 1984). This scale measures a thinking disposition of engaging in and enjoying effortful cognitive activity. The scale consists of statements such as “I would prefer complex to simple problems” and “I find satisfaction in deliberating hard and for long hours”. Participants are asked to indicate on a 5-point Likert scale to what degree the statements are characteristic of themselves. The end-points of the scale are 1 = “Extremely uncharacteristic of me” and 5 = “Extremely characteristic of me”. A total NFC score is calculated by adding up the score from all items, thus ranging from 18 – 90.

2.2.6 NASA task load index

The NASA task load index was used in Paper 1 and was included in studies 3 – 6. The N-TLX is a tool for measuring subjective task demand or mental workload associated with a task (Hart & Staveland, 1988). Workload is assessed on five separate scales. The scales use a visual analogue scale ranging from 1 = very low to 20 = very high. Participants indicate the experienced subjective mental demand, physical demand, temporal demand, performance, effort, and frustration. The scale was used to assess the cognitive effort associated with the other task paradigms in Paper 1.

2.2.7 Teleological reasoning task

The teleological reasoning task was included in Study 5: Day 1 and was the basis for Paper 2. The task was adapted from Kelemen et al. (2013) to be suitable for pupillometry. Teleological explanations refer to an endpoint or a final purpose (Kelemen et al., 2013). I.e., things exist for a purpose. An example of a teleological explanation is that “trees produce oxygen in order for people to breathe”. Although, common in religion, teleological explanations for natural phenomena are rejected in science where a mechanistic understanding of the universe prevail. However, teleological reasoning can be appropriate in the social-conventional and artifact domains. For example, “Schools exist in order to help people learn new things”. Importantly, humans tend to favor teleological explanations from an early age (Kelemen, 1999) and show a bias towards accepting false teleological explanations for natural phenomena in adulthood (Kelemen et al., 2013). That is, they show a teleological reasoning bias. In the teleological reasoning task participants are presented with 34 false teleological explanations for natural phenomena. These are test items and correct responding is judging them as false. Control items contain true and false physical explanations, and true and false teleological explanations. The control items containing true teleological explanations are in the social-conventional and artifact domains. The control items that are false teleological explanations are false due to incongruity, such as “Noses exist in order to support glasses”. There were 19 control teleological items. In addition, there were 24 control physical items that were true “Objects fall downwards because they are affected by gravity” and false “Rivers have rapids because a lot of fish swim in them”.

The Teleological reasoning task was computerized with stimuli presented auditorily. The task was self-paced and started by participants pressing the space bar on a keyboard. Start of the trials were indicated by the presentation of a fixation cross on screen. After approximately 500 ms (jittered) the stimulus sentences were presented auditorily. Stimulus sentences varied in length between 2.3 and 3.7 seconds. Participants responded by pressing “D” or “K” on the keyboard, indicating if the sentences were true or false, respectively. Responses had to be made after the end of the stimulus sentence presentation and within 4000ms after presentation. Feedback on responses was given after each trial, indicating if the answer was correct or incorrect, indicated by a “V” or “X”, respectively. Feedback was presented between 1800 – 2400 ms after responses. Duration of feedback presentation was between 4000 – 6200 ms.

2.2.8 Base-rate task

The underlying basis of the base-rate task was illustrated with an example “Tom W” in the section “1.2 Dual-process models of decision-making”. The base-rate task applied in Paper 3 in this thesis was adapted into two different versions to measure eye-gaze (gaze version) and pupil dilation (pupillometry version), respectively. The template for the two tasks was the base-rate task structure used in Pennycook et al. (2015b), which was adapted from De Neys & Glumicic (2008), who had adapted the tasks from Tversky & Kahneman (1973), see example “Tom W”. Data for the base-rate task was collected in Study 5: Day 2 and was the basis for Paper 3.

In the base-rate task from Paper 3, participants are asked to indicate which out of two groups (referred to as classes) a person most likely belongs to. Participants must decide based on two pieces of information. The first piece of information is the base-rate information. That is the composition of the two population groups, or number of people in each class. The base rates were extremely favoring one group (997 vs 3, 996 vs 4, 995 vs 5) or neutral (500 vs 500). The second piece of information was the personality trait or one-word attribute describing the person. This attribute was always congruent with a stereotype of one of the two classes. There were three conditions in both versions of the task, congruent, incongruent, and neutral. In the congruent condition both base-rate information and attribute favored the same class (20 trials). The correct response is the class favored by both types of information. In the incongruent condition the base-rate information and the attribute information favor opposing classes (40 trials). The correct response is the one favored by the base-rate information. In the neutral condition the base-rates are neutral, but the attribute favors one class (20 trials). The correct response is the class favored by the attribute.

Stimulus materials for the base-rate task were provided by Gordon Pennycook via personal correspondence. In this version of the base-rate task participants first receive on paper background information. The same as used in Pennycook et al. (2015b) was applied (see original article for details). Summarized, the background information describes that a large number of studies were carried out in a big research project. Every study contains two population groups (e.g., nannies and lawyers). Participants will receive information regarding the composition of the two population groups. Further, a person is drawn at random from each study. This person is described by a one-word personality trait or attribute. The participants' task is to decide which of the two groups the person most likely was drawn from. Following the background information, participants were presented with a power point presentation familiarizing participants with the task structure. Additionally, participants conducted three practice trials before starting the experiment.

In the gaze version of the task, participants imitate each trial by pressing the spacebar. After 500 ms, a fixation cross is presented for 500ms. Followed by the information "this study contains", and the class information (two population groups), presented on screen for 1800ms. Then, a fixation cross appears for 200 ms, followed by the attribute describing the person, presented for 1800 ms. Followed by a fixation cross, presented for 200 ms. The response slide is presented for maximally 4000 ms or terminates when a response is made. On the response slide the class information is presented on top and the base-rate associated with each class is presented directly underneath. The response slide is followed by a 200 ms fixation cross, then 1800 ms of a blank screen. Participants indicate their choice, left or right option (options divided vertically on screen) by pressing "A" or "L", respectively.

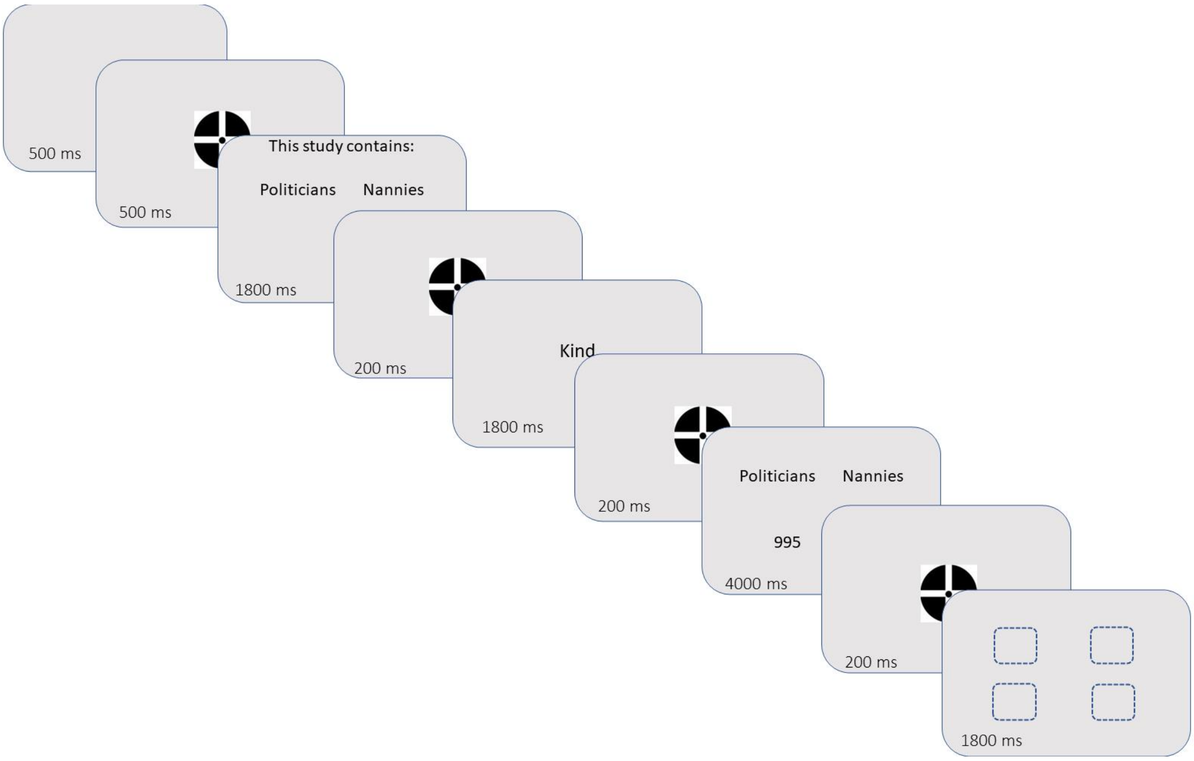


Figure 3. Trial structure in Base-rate – gaze version.

Note. The figure illustrates temporal order of a trial. Illustrations do not accurately represent stimuli presented on screen. The dotted squares mark the areas of interest for gaze analysis and are not visible on the screen.

In the pupillometry version of the task, information is split between being presented on screen and auditorily via noise cancelling headphones. See figure 4 for an illustration of trial structure.

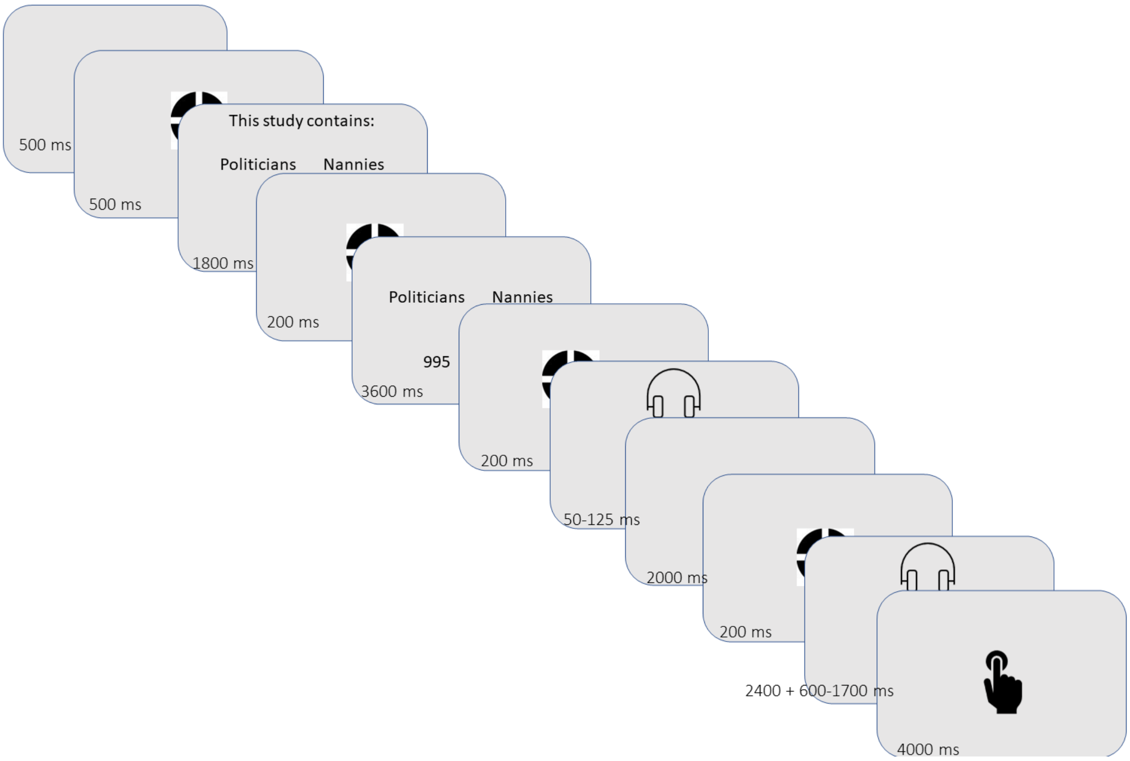


Figure 4. Temporal structure of the base-rate task - pupillometry version.

Note. The figure illustrates the structure of a trial in the base-rate task – pupillometry version. Illustrations do not accurately represent stimuli presented on screen. Headphones indicate auditory stimulus.

The task is initiated by participants pressing the space bar for each trial. Followed by 200 ms fixation cross. Then the text “This study contains:” is presented with the class information for 1800 ms. This is followed by 200 ms fixation cross, and subsequently both class information (presented side by side top part of screen) and base-rate information (presented below the corresponding class) was presented for 3600 ms, followed 200 ms fixation cross. Then a blank screen was shown while the attribute information was presented auditorily. Length of attribute sound files were between 50 – 120 ms. The attribute sound file was followed by 2000 ms of continued blank screen. Then, a fixation cross was presented for 200 ms, followed auditory presentation of a question “is this person more likely a…”, directly followed by auditory presentation of a class (example “politician”). The length of the question audio file was 2400 ms, and class audio file varied between 600 ms and 1700 ms. After both the question and class audio files participants had 400 0ms to respond. Pressing “A” to indicate yes, or “L” to indicate no. Labels behind the keyboard indicated which response were associated with yes and no responses to avoid confusion and additional mental load.

2.3 Pupillometry

In Study 5 (Day 1 and Day 2) pupillometry was an essential measure. Therefore, the following exclusion criteria were applied. Participants had to report that they had no previous history of brain disease, neurological disorder or brain surgery. Participants could not be taking any medications or drugs that affects the central nervous system. Additionally, the stimulus materials were presented in English. Therefore, participants had to rate their English proficiency on a 7-point Likert scale where the end points were 1 = "understand a few words" and 7 = "master it like native language". Participants who rated their English proficiency lower than 4 (not including 4) were excluded from participation. This criterion was set based on a previous study which found no differences in rational reasoning due to language for participants scoring higher than 4 on the same scale (Mækelæ & Pfuhl, 2019).

For Study 5 (Day 1 and Day 2) a video-based infrared eye-tracker, the Eyelink 1000 (SR Research) was used for measuring eye-gaze in the base-rate gaze version, and for measuring pupil size in both the teleological reasoning task and the base-rate pupillometry version. Recordings were sampled at a rate of 500 Hz. Artifacts caused by eyeblinks, head-movements etc. were detected based on the velocity of the signal (Mathôt et al., 2018). Artifacts were corrected with linear interpolation. Thresholds for interpolation were adapted on an individual basis. The signal was smoothed with a 3 Hz low pass Butterworth filter. The signal was treated as missing for artifacts lasting more than 1000 milliseconds. The signal was visually screened and trials with artifacts remaining in the time-windows of interest were excluded. Time-windows of interest where the interpolated signal was missing for more than 50% of the time were treated as missing. Before each pupillometry task a 2-minute baseline pupil measure was recorded (not included in analyses).

For Paper 2, pupil size was recorded during the teleological reasoning task. The time-windows of interest were trial-baseline pupil size, pupil dilation before decision, and pupil dilation to feedback. The trial-baseline pupil size was recorded as the average signal in the 200 ms following fixation cross onset at the start of each trial. Indicating fluctuating levels of LC-NE activity and neural gain. Also used for baseline correction of pupil dilation measures. Pupil dilation before decision was measured as the maximum dilation from stimulus onset until a response was made. Pupil dilation before decision was baseline corrected by subtraction (for exploratory measures the non-corrected signal was also applied). The measure indicated cognitive effort, or a reverse measure of neural gain (predicted by dual-process theory and extensive integration, respectively). Pupil dilation to feedback was recorded as the maximum recorded pupil size in the time window from feedback onset and the following 3000 ms. The signal was used to indicate surprise and uncertainty. Pupil dilation to feedback is not discussed further in this thesis. Baseline pupil size and pupil dilation measures were z-scored within-participants.

Paper 3, the base-rate task, gaze version. Four quadrants surrounding the class and base-rate information were pre-defined as areas-of interest (AOI). The gaze measure of interest was total gaze duration inside the AOI's, recorded during presentation of the response slide (both class and base-rate information presented). Proportional gaze was calculated for type of information (class or base-rate).

This was calculated as gaze at the two top quadrants surrounding the class information, divided by total gaze inside all four AOI's. Proportional gaze at correct option was recorded, but not further discussed in this thesis. Proportional gaze at correct option was calculated as total gaze time inside the two AOI's (top and bottom) on the side of the correct response (vertically divided), left or right, divided by total gaze inside all four AOI's.

For Paper 3, pupil size was recorded for the base-rate task pupillometry version. The time-windows of interest were trial-baseline pupil size, pupil dilation following attribute information (attribute time window) and pupil dilation before decision (decision time-window). Pre-processing for the task revealed data loss in recordings. Participants with less than 40% valid trials in the congruent and incongruent conditions were excluded separately for further analyses in the time-windows of interest. The trial-baseline pupil size (N = 47) was measured as the average signal in the 200 ms following fixation cross onset at the start of each trial. This measure was used for baseline-correction and exploratory investigation of alternative hypotheses (attention, mind-wandering, adaptive-gain theory). Pupil dilation in the attribute time-window (N = 37) was measured as the maximum pupil size recorded in the time period from the end of the attribute sound file and the following 2000 ms. The measure was used to indicate conflict detection and cognitive effort (and also indicates LC activity). The pupil dilation before decision (N = 38) was recorded from the end of the question sound file to a response was made, maximum 4000 ms. The measures were used to indicate cognitive decoupling and cognitive effort (and also indicate LC activity). Both pupil dilation measures were baseline corrected by subtracting the baseline pupil size, and z-scored within participants.

2.4 Data analysis

Variables were assessed for normality and outliers by Shapiro-Wilks test and visual inspection of box plots, frequency distributions and QQ-plots. Non-parametric tests were applied when assumptions of parametric tests were not met. Statistical tests were conducted in R using RStudio. Mixed models were analyzed with the lme4 package (Bates et al., 2015). Residuals were inspected with the DHARMa package (Hartig, 2022) and variance inflation factor with the caret package (Kuhn, 2015). Drift-diffusion modelling of responses was performed with Python (Patil et al., 2010) version 3.9 using the python toolbox HDDM (Wiecki et al., 2013). In Paper 3, a dockerHDDM was used (Pan et al., 2022). The DDM's for Paper 2 and Paper 3 were run with five Markov chains with 20,000 samples, 12,000 burn-in, and every second sample discarded as thinning. Model convergence was assessed with the Gelman-Rubin R statistic and visual inspection of the trace, autocorrelation and marginal posterior. The deviance information criterion was used for model comparison.

In Paper 1, bivariate Pearson correlations between cognitive effort tasks were first calculated for each study separately. The metafor package (Viechtbauer, 2010) was then used to calculate mean effect sizes based on the correlation coefficients from each study. Linear mixed models were conducted separately for COG-ED, DST and rational reasoning battery, to assess if any of the measures could be predicted by NFC, perceived effort in N-TLX, working memory capacity, or any of the other behavioral cognitive effort measures.

In Paper 2, generalized mixed models were applied for analyzing response times, pupil dilation and baseline pupil size. Drift-diffusion modelling of responses was applied in two stages. First, the main parameters of the model were estimated and tested for differences between the test- and control condition. Second, the pupil measures were entered as predictors of trial-by-trial variation in drift-rate, threshold and drift-rate variability.

In Paper 3, for both base-rate tasks, participants with accuracy rates three standard deviations below average in the congruent and neutral conditions were excluded. Response times faster than 150 ms and slower than 4000 ms were excluded. As accuracy rates in the incongruent condition revealed a bi-modal distribution, the data was exploratorily analyzed separately for two distinct groups of responders. The gaze version and the pupillometry version were analyzed separately. In the gaze version of the base-rate task proportional gaze at information type and correct choice option was investigated, in addition to accuracy and response times. In the pupillometry version of the base-rate task, pupil dilation following attribute (conflict detection) and before decisions (cognitive decoupling) were analyzed, in addition to trial baseline pupil size, response times and accuracy. The drift-diffusion model was analyzed separately for the two base-rate tasks, investigating differences in the main parameters for the three conditions.

2.4.1 Sample size

Sample size minimum for Paper 1 (N = 402) allowed to detect correlations of r = .177 and higher at an alpha level of 0.05 and with power of .95.

For study 5, a power calculation was conducted based on a previous study (de Berker et al., 2016) finding a pupil measure of uncertainty correlated with performance (effect size r = 0.62, N=22). Assuming regression to the mean, the calculation was based on a smaller effect size of r = .4. In addition, we used an alpha level of 0.05 (two-sided) and power of 0.8. This resulted in a sample size of 44 participants. For Paper 2, regarding individual differences large effect sizes have been found with partial η^2 between 0.3 and 0.6 (Thompson et al., 2018). A sample of 40 participants was considered sufficient to find an effect.

3 Summary of papers

3.1 Paper 1)

3.1.1 Aims and background

Cognitive effort is ubiquitous, however the measurement of cognitive effort remains a challenge (Thomson & Oppenheimer, 2022). Cognitive effort is a function of both cognitive ability and motivation to perform the task at hand (Shenhav, Musslick, et al., 2017; Westbrook & Braver, 2015). Trait differences in cognitive motivation can be reliably measured with the need for cognition (NFC) scale (Cacioppo & Petty, 1982; Hussey & Hughes, 2020). In recent years multiple behavioral tools to measure cognitive effort have been developed to measure actualized cognitive effort expenditure (Frederick, 2005; Kool et al., 2010; Stanovich, 2016; Toplak et al., 2011; Westbrook et al., 2013). This includes the demand selection task (DST), the cognitive effort discounting paradigm (COG-ED), and rational reasoning task batteries, applying tasks from the heuristics and bias literature. Further, concerns about the reliability and validity of self-report motivate the use of behavioral task paradigms (Paulhus & Vazire, 2007). As these behavioral tools are being widely applied and used interchangeably there is a need to assess to what extent these tools are related to each other and evaluate if they are measuring the same cognitive effort construct (Chang et al., 2020; Culbreth et al., 2016; J. M. Gold et al., 2015; Puveendrakumaran et al., 2020; Thomson & Oppenheimer, 2022). The aim of this paper was to assess the shared variance of three cognitive effort measures, DST, COG-ED and rational reasoning battery, and their relationship to NFC. Additionally, investigating test-retest reliability of COG-ED, DST, and rational reasoning battery and the relation to working memory and subjective perceived mental effort as assessed by the NASA task load index (N-TLX; Hart & Staveland, 1988).

3.1.2 Methods

Six studies were conducted by two independent labs with a total of 663 participants. Study 1 and 2 were conducted at Washington University in St. Louis, USA. Study 3 – 6 were conducted at UiT–The Arctic University of Norway. All studies included the NFC, and two or three of the behavioral measures of cognitive effort. Only three of the studies included the N-TLX. In the COG-ED (Westbrook et al., 2013) the n-back task was played for six load levels (N = 1 - 6) in Study 1 and 2, and for four levels (N = 1 - 4) in Study 3 – 6. The average subjective indifference point (AIP) across all load levels was used as the cognitive effort measure in COG-ED. Working memory capacity was measured with the d' (calculated using signal detection theory) from the n-back task in COG-ED. The DST was applied in Study 1 - 5 with an exact replication of study 3 in Kool et al. (2010), with cognitive demand avoidance (scored as proportion of high demand choices, 0 – 0.5 indicates demand avoidance, 0.5 – 1 indicates demand preference) as the measure of cognitive effort. In Study 6 we used the abridged version (Patzelt et al., 2019). Rational reasoning was measured with a combination of items from the heuristics and bias literature (see Toplak et al., 2011), including the cognitive reflection test (Toplak et al., 2014), number of items varied between 12 – 18. Proportion of correct items (no items correct = 0, all items correct = 1) was calculated as the cognitive effort measure. Perceived effort was measured with the N-TLX on a visual analogue scale (1 = very low, 20 = very high) (Hart & Staveland, 1988). NFC was measured with the 18-item scale in all six studies (Cacioppo et al., 1984). A meta-analytic approach was applied to analyze the overall effect sizes. We conducted Pearson's correlations for each study. Based on the effect size from each study we calculated an overall correlation based on the effect size and sample size of each study by using the metafor package (Viechtbauer, 2010).

3.1.3 Results

The results show a significant positive association between NFC and AIP in COG-ED. Meaning higher NFC is related to less discounting of monetary rewards due to cognitive effort. Less cognitive effort discounting in the COG-ED was also related to higher working memory capacity. The results show a significant positive association between NFC and rational reasoning. Indicating that individuals higher in NFC tend to perform better on rational reasoning or heuristics and bias tasks. Higher rational reasoning score was also associated with higher working memory capacity and rating the task as less effortful. Cognitive effort discounting in COG-ED was not related to rational reasoning performance. Additionally, demand avoidance in the DST was not related to any other effort measure, nor working memory or perceived effort in N-TLX.

3.1.4 Conclusion

Cognitive effort is a difficult concept to measure. We find no association between three behavioral measures of cognitive effort. However, both cognitive effort discounting in COG-ED and rational reasoning performance were related to NFC and working memory capacity. This study shows that results found using one of these cognitive effort measures cannot be assumed to apply to other cognitive effort measures. We conclude that the COG-ED paradigm is a valid cognitive effort measure, but relies on external rewards, which can be a confounder. Further, the results suggests that the rational reasoning battery and the DST are not applicable individual difference measures of cognitive effort. Our study highlights the need to develop new behavioral tools for measuring cognitive effort. There is an absence of a reliable behavioral measure of internal motivation for cognitive effort expenditure.

3.2 Paper 2)

3.2.1 Aims and background

Human decision-making is prone to bias. An example of such a bias is the tendency for humans to see purpose and intentionality in phenomena when there is none, that is, a teleological reasoning bias (Kelemen et al., 2013). Competing decision-making frameworks propose opposing explanations for how bias in reasoning occur (Eldar et al., 2021). The default-interventionist, dual-process account proposes that errors in reasoning occur due to the fallibility of fast effortless Type 1 reasoning processes, and a failure to engage slower more effortful deliberate Type 2 reasoning process (J. St. B. T. Evans & Stanovich, 2013; Kahneman, 2011). An alternative dual-process account, the smart intuitor, proposes that smarter individuals have better intuitions compared to individuals of lower intelligence (Raoelison et al., 2020). This account maintains that in many cases smarter individuals will not make reasoning errors because they have correct intuitions. This account builds upon findings that most of correct responses on heuristics and bias tasks result from fast effortless Type 1 processes, and only a small portion of correct responses occur due to slower, effortful deliberate Type 2 processing (Raoelison et al., 2020). However, the account still maintains that slower, effortful, deliberate Type 2 processing should more likely lead to correct responses, whereas fast, effortless Type 1 processing should result in more biased responses. Importantly, larger pupil dilations are associated with more cognitive effort, which can therefore be used as a measure of deliberate Type 2 processing (Kahneman & Beatty, 1966; van der Wel & van Steenbergen, 2018). On the contrary, the extensive integration account proposes that bias in reasoning is exacerbated by more extensive evidence accumulation in sequential sampling models (Eldar et al., 2021). The extensive integration account builds upon a framework where the decision-making process is seen as a gradual noisy evidence accumulation process towards some alternatives, until a decision threshold is reached, triggering a decision (Busemeyer et al., 2006; Krajbich & Rangel, 2011; Usher et al., 2013). According to this account, low neural gain is associated with less evidence weighting (Aston-Jones & Cohen, 2005; Usher et al., 1999). Thus, requiring more extensive evidence accumulation, leading to more bias (Eldar et al., 2021). Neural gain can be assessed using pupillometry as the pupil is highly correlated with Locus Coeruleus activity and the Locus Coeruleus - Norepinephrine system modulates neural gain (Aston-Jones & Cohen, 2005; Eldar et al., 2013; Gilzenrat et al., 2010; Reimer et al., 2016). Low neural gain is associated with smaller baseline pupil size and larger pupil dilations, since baseline pupil size and baseline-corrected pupil dilations are negatively correlated (Aston-Jones & Cohen, 2005; Eldar et al., 2013; Gilzenrat et al., 2010). Thus, the extensive integration account predicts that bias in reasoning is associated with larger pupil dilations, indicating low neural gain, and longer response times. On the contrary dual-processing accounts propose that bias in reasoning is more likely to occur due to fast effortless (Type 1) processing, indicated by smaller pupil dilations and faster response times. The aim of this study was to test predictions of competing decision-making frameworks explaining bias in reasoning. This is done by testing if bias in reasoning on a teleological reasoning task is associated with more or less cognitive effort as measured with pupillometry, and if bias is associated with slower or faster response times. Additionally, we included individual difference measures of cognitive ability and cognitive motivation to dissociate predictions from the two dual-process accounts.

3.2.2 Methods

Participants (N = 45) performed a teleological reasoning task (Kelemen et al., 2013). The task had been computerized and adapted to be suitable for pupillometry with stimulus sentences being presented auditorily. Responses were made on a keyboard to signal if the stimulus sentences were true or false. The task consisted of false teleological explanations for natural phenomena, referred to as test items (34 items), and control statements consisting of true and false physical explanations (24 items) as well as true and false teleological explanations in the socio-cultural domain (19 items), where these explanations are appropriate. Additionally, participants completed measures of cognitive motivation (NFC) and cognitive ability (rational reasoning items from the heuristics and bias literature). Primary outcome measures were accuracy rates, response times, and baseline-corrected pupil size on the teleological reasoning task. Across participants we looked at individual difference measures. As exploratory analyzes we looked separately at baseline pupil size and pupil dilation (non-corrected). We analyzed the response times and the pupil measures with a drift-diffusion model to further test predictions from the extensive integration account. Lastly, we included measures of pupil dilation to feedback to assess uncertainty and surprise.

3.2.3 Results

We replicate that individuals are prone to accept false teleological explanations for natural phenomena and show a teleological reasoning bias. We find that bias in reasoning as measured by errors on the teleological reasoning task is associated with slower response times, smaller baseline pupil size and larger pupil dilations. Thus, the results are in line with the extensive integration account of bias in reasoning and show results directly opposing predictions from dual-process theories. We find that performance on the teleological reasoning task is related to cognitive ability but not cognitive motivation. Additionally, by modelling responses on the teleological reasoning task with a drift diffusion model we show that larger baseline pupil size is associated with lower decision threshold and higher drift-rate. Conversely, larger pupil dilations are associated with lower drift-rate and higher decision-threshold. Thus, following the predicted relationship from the extensive integration account between pupil size and evidence accumulation, modulated by neural gain.

3.2.4 Conclusion

Bias in teleological reasoning is associated with larger pupil dilations and longer response times. This supports sequential sampling models and the extensive integration account of bias in reasoning. It highlights the role of the Locus-Coeruleus – Norepinephrine system in decision-making and bias in reasoning. The results from the study directly oppose predictions from dual-processing accounts, where biases in reasoning result from fast effortless processing.

3.3 Paper 3)

3.3.1 Aims and background

The study of reasoning errors has been highly influential in the development of dual-process theories of reasoning and decision making (J. St. B. T. Evans, 2008; Kahneman, 2011; Kahneman & Frederick, 2002; Kahneman & Tversky, 1973). However, an accumulating body of empirical research finds results opposing predictions from classical dual-process theories (Bago & De Neys, 2019; Newman et al., 2017; Raoelison et al., 2020). This has led to a new generation of dual-process models of decision making (De Neys & Pennycook, 2019; Pennycook et al., 2015b; Raoelison et al., 2020). Pennycook et al. (Pennycook et al., 2015b) postulate a three-stage model of analytic engagement. In this model competing intuitions give rise to deliberate, slow, effortful Type 2 reasoning. Failure to detect a conflict between competing intuitions and rationalization of an incorrect response are considered early and late sources of bias in this model. Successful reasoning performance is therefore dependent on conflict detection and cognitive decoupling (successful Type 2 reasoning). The model is validated by Pennycook et al. (2015) by measuring response times on an adapted version of the classical base-rate neglect task (De Neys & Glumicic, 2008; Kahneman & Tversky, 1973). In the base-rate task participants are given two pieces of information which should guide them to decide which out of two groups a person drawn at random most likely belongs to. The two pieces of information to guide their choice are 1) a one-word attribute describing the person, and 2) the number of individuals in each group, i.e., the base-rate (Pennycook et al., 2015b). In this study, we adapted the base-rate neglect task and created two versions of the task to be compatible with eye-tracking and pupillometry, respectively. Our aim was to test the three-stage model of analytic engagement (Pennycook et al., 2015b), while simultaneously investigating visual attention (eye-tracking) and cognitive effort (pupillometry) in a classical decision-making task.

3.3.2 Methods

A total of 60 participants took part in the study. Participants completed two versions of the base-rate task. One version (gaze version) of the base-rate task was adapted for the use of eye-tracing to measure eye-gaze during performance as an indicator of visual attention. In this task the attribute information was displayed before the base-rate information. The second version (pupillometry version) of the base-rate task was adapted for pupillometry to measure pupil size during task performance as an indicator of cognitive effort. In this version of the task the base-rate information was presented before the attribute information. The tasks were counterbalanced across participants. Individual difference measures of cognitive ability, cognitive motivation and personality traits were of secondary importance and performed before and in between the two base-rate tasks. The outcome measures of interest were choices and response times. Additionally, in the gaze version the proportion of gaze at the two pieces of information, and the proportion of gaze at the correct option were of interest. In the pupillometry version of the task the pupil measures of interest were pupil size at trial baseline, after presentation of the attribute information (i.e., in the time window of conflict detection), and before making a decision (i.e., measuring cognitive decoupling). We applied a drift diffusion model to responses in both base-rate tasks to analyze differences in decision parameters in the three conditions, specifically we wanted to test if conflicting information led to increased decision threshold (Lin et al., 2023). Lastly, as exploratory analyses we separated participants based on their majority response either base-rate congruent or stereotype congruent, as we found evidence for a bi-modal distribution of responses across participants.

3.3.3 Results

The results from both base-rate tasks revealed that participants could be separated into two groups based on their responses in the incongruent condition. One group gave the stereotype congruent response on the majority of trials ("stereotype responders") and the other group gave the base-rate congruent response on the majority of trials ("base-rate responders"). Analyses of response times, eye gaze and pupil dilation across conditions on the two tasks revealed that there were no significant differences for the stereotype responders. The base-rate responders on the other hand showed significant differences in response time, eye-gaze and pupil dilation across conditions. This indicates that conflict detection failure was prominent among stereotype responders, and base-rate neglect is an important source of biased responses on this task. When analyzing responses in the incongruent condition (either base-rate congruent or stereotype congruent answers), the two groups were faster when giving their majority response (i.e., stereotype congruent response for the stereotype responders), compared to when giving the minority response (i.e., base-rate congruent response for the stereotype responders). Additionally, response times for the two response options in the incongruent condition (stereotype congruent or base-rate congruent) was opposite in the two tasks. That is, base-rate congruent responses were slower in the incongruent condition in the gaze version of the task, whereas the same response was faster in the pupillometry version. A result that was further supported by a drift diffusion model showing opposing bias in the two tasks (congruent with the aforementioned response times). Analyzes with the drift diffusion model across conditions further showed slower drift-rate in the incongruent condition, not increased decision threshold, indicating increased task difficulty rather than increased deliberation. In the gaze version of the task base-rate responders tended to look more at the base-rate information compared to stereotype responders. However, both groups tended to look more at the class (group) information compared to the base-rate information, and more at the option they ended up choosing.

In the pupillometry version of the task, smaller baseline pupil size and larger pupil dilations before decisions were related to performance across conditions. When comparing pupil size across conditions in the attribute time window we found no significant difference in pupil dilations for the sample as a whole. When looking separately at the groups of responders there was no significant difference for the stereotype responders. However, for the base-rate responders pupil dilations in the attribute time window were largest in the neutral condition i.e., when the base-rate information was not informative. Similarly, before making a decision, pupil dilations were largest in the neutral condition. This effect was not significant for the stereotype responders. However, for the base-rate responders both the neutral and incongruent condition were associated with larger pupil dilations compared to the congruent condition. Accordingly, base-rate responders use more effort before deciding when there is conflicting information or when they can no longer use the base-rate information.

When comparing responses in the incongruent condition in the attribute time window, stereotype responders showed larger pupil dilations before making a correct response. For the base-rate responders there was no difference between trials where they make errors or correct responses. For the sample as a whole pupil dilation was not a significant predictor of correct responses in the attribute time window. When comparing responses in the incongruent condition before making a decision, pupil size was not a significant predictor of responses.

3.3.4 Conclusion

By applying two versions of a well-established reasoning task adapted for eye-tracking and pupillometry, respectively, we evaluated the three-stage model of analytic engagement. By analyzing two groups of participants separately we find one group which shows the expected base-rate neglect, and support for conflict detection failure. The other group is sensitive to changes in base-rate information, as seen by changes in response times, eye-gaze and pupil dilation. When analyzing these two groups separately, we show that the expected slowing of responses for base-rate congruent responses in the incongruent condition can be reversed and is dependent on task manipulations and participants majority response, suggesting a gap in the existing literature. By measuring pupil dilation, we find preliminary support for the constructs of conflict detection and cognitive decoupling. The results suggest that phasic LC activity and cognitive effort may be implicated in changing responses from stereotype congruent to base-rate congruent. This is consistent with phasic LC activity acting as a neural interrupt signal and perhaps being involved in conflict detection. Further, when comparing conditions, we found larger pupil dilations for the neutral condition compared to the congruent condition in both the attribute and decision time window for the base-rate responders. Additionally, the incongruent condition was associated with larger pupil dilations compared to the congruent condition before making a decision for the base-rate responders. Thus, we find evidence of effort differences across conditions, contrary to predictions of the three-stage model the neutral condition was associated with the most effort. Note that there was some loss of data due to noise in the measurement of pupil dilation. Thus, due to sample size the results regarding pupil dilation should be considered preliminary.

4 Discussion

The primary aim of this thesis was to investigate the role of cognitive effort in decision-making and reasoning errors. In order to do this, we had secondary aims to evaluate tools measuring cognitive effort and compare decision-making frameworks explaining errors and bias in reasoning. It has been a long-standing proposition that errors in reasoning occur due to fast, effortless reasoning and that more cognitive effort and deliberation will lead to better decisions. In this thesis this hypothesis is explored through several approaches. Furthermore, the development of new behavioral tools investigating cognitive effort has significantly advanced our understanding of cognitive effort. However, there is little data comparing these measurement tools with each other to assess their shared variance. This thesis compares multiple tools of cognitive effort to elucidate the strength and weaknesses of these tools and their shared variance. Lastly, accumulating evidence contradicts predictions from the long-standing influential default-interventionist, dual-process theory of decision-making. This has inspired the development of new decision-making paradigms. In this thesis I test predictions from different decision-making paradigms to evaluate them.

4.1 The role of cognitive effort in reasoning errors

The main findings concerning the role of cognitive effort in reasoning errors is, in Paper 1, a battery of rational reasoning tasks, was found to not be related to other behavioral measures of cognitive effort, namely, DST and COG-ED. Additionally, rational reasoning performance was negatively related to subjective effort experienced on the task as measured with the N-TLX. However, a thinking disposition of enjoying cognitively demanding activity (NFC) was related to higher performance on the rational reasoning battery. In Paper 2, I found that errors on a teleological reasoning task was related to larger pupil dilations, indicating more cognitive effort. However, in Paper 3 I found that larger pupil dilations were associated with base-rate congruent (correct) responding in the incongruent condition on the base-rate task. Additionally, when receiving conflicting information in the attribute time window, larger pupil dilations were associated with changing responses from stereotype congruent, i.e., making a reasoning error, to the correct base-rate congruent response, for those tending to make a base-rate neglect reasoning error.

The finding that there was no relationship between rational reasoning tasks and other behavioral measures of cognitive effort indicates that lower individual cost of cognitive effort, or effort discounting of rewards, and cognitive demand avoidance cannot predict if individuals are prone to make reasoning errors. These results indicate no association between cognitive demand avoidance or high cognitive effort costs being related to reasoning errors, and thus no relation between cognitive effort and reasoning errors. This study is to my knowledge the first study to assess the relationship between rational reasoning tasks and these behavioral cognitive effort measures. The result that there is no relationship between these measures is contrary to the long-held assumption of cognitive effort being linked to errors in reasoning (J. St. B. T. Evans, 2008; Kahneman, 2011). Furthermore, I find that the subjective mental effort experienced on the rational reasoning tasks, which I assume is an indicator of expended cognitive effort, is negatively related to performance on these tasks. That is, participants who expended more cognitive effort on the rational reasoning tasks performed worse. This finding also goes against the proposal that higher cognitive effort leads to overcoming reasoning errors (Frederick, 2005; Kahneman & Frederick, 2002). Importantly, this finding could be explained by participants with lower performance on the rational reasoning tasks spending more time and effort on the tasks due to their inability to solve the tasks (or solve them quickly) compared to higher performing individuals who solve the tasks quickly and with little or less effort. This explanation is consistent with the smart intuitor account which proposes that higher cognitive ability individuals solve these tasks quickly, whereas few participants, who initially fail to solve the tasks, manage to correctly solve them when given more time (Bago & De Neys, 2017; Newman et al., 2017; Raoelison et al., 2020). The finding that performance on rational reasoning tasks was related to a thinking disposition of enjoying cognitively demanding activity, on the contrary suggests a positive link between cognitive effort and overcoming reasoning errors. It might be that individuals who enjoy cognitive challenges perform better on the rational reasoning tasks, and that the enjoyment of complex challenges measures something different than the cognitive effort paradigms which use repetitive tasks such as used in the DST and COG-ED. NFC might capture something more than what is captured with DST and COGED. A positive relationship between NFC and performance on heuristics and bias tasks is consistent with previous studies (Thomson & Oppenheimer, 2016; Toplak et al., 2014; West et al., 2008) and will be discussed further, under the section 4.2 Comparing tools measuring cognitive effort. Overall, the result from Paper 1 suggests that cognitive effort is not related to errors in reasoning, in fact it suggests that participants who make more errors in reasoning invest more effort on rational reasoning tasks. However, NFC may measure something that is related to avoiding reasoning errors.

Teleological reasoning bias was found to be associated with larger pupil dilations, indicating more cognitive effort, in Paper 2. This suggests that this error in reasoning is associated with more not less cognitive effort. This is contrary to the proposal from Kelemen et al. (Kelemen et al., 2013) that teleological reasoning remains a life-long cognitive default in understanding natural phenomena that have to be overridden, an effortful process, in order to produce a scientific/mechanistic understanding. However, this result is consistent with Eldar et al. (Eldar et al., 2021) who found that larger pupil dilations were related to more reasoning bias on three framing tasks. Interestingly, smaller baseline pupil size was also related to errors in reasoning on the teleological reasoning tasks. This suggests that a certain level of attention, alertness, arousal, and perhaps a minimum level of cognitive effort, need to be present for optimal task performance. This is consistent with research on attention and the LC-NE system (Aston-Jones & Cohen, 2005; Unsworth & Robison, 2016). Alternatively, as suggested by the extensive integration account, low neural gain leads to more extensive information sampling where bias accumulates, leading to more biased responses (Eldar et al., 2021). An alternative explanation for these results is that participants invested more effort when they experienced more uncertainty or failed to come up with the correct answer, similar to the finding of higher subjective effort being related to worse performance on rational reasoning battery in Paper 1. However, there were no difference in effort between the test and control condition which speaks against this explanation. Overall, the results from Paper 2 suggests that errors in teleological reasoning are associated with more not less cognitive effort, which is partly contrary to the findings from Paper 3.

In the base-rate task, the results indicated that neglecting the base-rate information was a significant source of bias on this task for a group of participants, the stereotype responders. Comparing pupil dilation in the congruent (non-conflicting information) and the incongruent (conflicting information) conditions. I found that base-rate responders showed larger pupil dilations, i.e., more effort, when there was conflicting information, whereas stereotype responders showed no difference between conflicting and non-conflicting information. Evidence of neglecting base-rate information for stereotype responders could also be found in response time differences and eye-gaze differences between conditions. Thus, one group of participants made significantly more reasoning errors on this task by neglecting the base-rate information, and in doing so, spent less cognitive effort. This finding suggests that less cognitive effort is related to reasoning errors on this task. Additionally, stereotype-responders showed larger pupil dilations when giving the correct (base-rate congruent) response in the incongruent condition, further supporting cognitive effort being associated with not making the base-rate neglect reasoning error on this task. For the stereotype responders it might be that overriding or inhibiting their usual response required more effort. For the base-rate responders it is probable that errors on the task are related to attentional deficits when not giving the correct response as they on average performed well on the task. This would be consistent with studies suggesting that inhibition and updating requires cognitive effort (Friedman & Miyake, 2017; van der Wel & van Steenbergen, 2018) and that lapses of attention are related to smaller phasic pupil dilations (Unsworth & Robison, 2016). Furthermore, that stereotype responders showed larger pupil dilations after receiving conflicting information (incongruent condition) on trials where they made the correct response, which was opposite to their majority response is consistent with theories proposing that phasic LC activity acts as a “neural interrupt signal” or reorienting (Bouret & Sara, 2005; Dayan & Yu, 2006) and the LC-NE system being involved in cognitive flexibility (McGaughy et al., 2008). It should be noted that when comparing the three task conditions for the whole sample there was only a difference in pupil dilations in the attribute time-window not before making a decision. Importantly, I found that larger pupil dilations before decisions were associated with higher performance across conditions. Indicating that larger pupil dilations may be associated with general performance (Aston-Jones & Cohen, 2005; Unsworth & Robison, 2016; van der Wel & van Steenbergen, 2018). Overall, the results from Paper 3 suggests that cognitive effort is involved in correct responses on the base-rate task, and that a lack of cognitive effort is associated with reasoning errors on this task.

The results regarding pupil dilation for Paper 2 and Paper 3 show partly opposing results. A central difference between these tasks is that in the base-rate task there are multiple pieces of information which can be conflicting, or not, and which can be ignored or not. This is in contrast to the teleological reasoning task where there was no conflicting information, and no obvious strategy to reduce the amount of information.[3] In the base-rate task, performance was dependent on integrating several pieces of information, a process similar to the executive function updating. Additionally, some participants had to inhibit responses. These executive control actions require cognitive effort which is reflected in larger pupil dilations (Friedman & Miyake, 2017; van der Wel & van Steenbergen, 2018). Thus, in the base-rate task there was a need for cognitive effort due to the executive functions of updating and inhibition. Additionally, participants could ignore some cues and integrate less information (Shah & Oppenheimer, 2008). This is in contrast to the teleological reasoning task where there were no obvious effort reducing strategies available and no difference in executive control demand for correct and error responses.

Regarding the question of the role of cognitive effort in reasoning errors. A composite of common reasoning errors was not related to cognitive effort avoidance or cognitive effort discounting. Subjective cognitive effort ratings suggested that more effort was associated with worse performance. On a teleological reasoning task more cognitive effort was associated with reasoning errors. These findings indicate that more cognitive effort is not associated with overcoming reasoning errors, and more cognitive effort can be associated with worse performance. More cognitive effort being associated with errors in reasoning is consistent with Eldar et al. (Eldar et al., 2021), and compatible with proposals such as rationalization being the main operating mode of Type 2 processing (De Neys, 2020; J. St. B. T. Evans, 2019) and accounts of motivated reasoning (Kahan, 2013; Kunda, 1990; Pennycook & Rand, 2019; Persson et al., 2021). On the contrary, I did find that people who enjoy cognitively demanding activity performed better on the reasoning tasks and that more cognitive effort was related to better performance on the base-rate task. It is plausible that NFC assesses something more than a tendency to expend cognitive effort, which might be related to intelligence and enjoyment of mental challenges. The finding that cognitive effort is related to performance on the base-rate task suggests that for some reasoning errors a lack of cognitive effort, either in not expending enough cognitive effort or not recognizing the need to engage cognitive effort (or integrate information), is a central factor. This is consistent with the proposed relationship between cognitive effort and reasoning errors put forth by dual-process proponents (J. St. B. T. Evans, 2008; Frederick, 2005; Kahneman, 2011; Kahneman & Frederick, 2002). To summarize, some reasoning errors might be caused by a lack of cognitive effort, however some reasoning errors might be more prominent with more cognitive effort. Additionally, more cognitive effort can be linked to worse performance on reasoning tasks. When reasoning problems such as used in the heuristics and bias literature are intermixed, they show no relation to behavioral measures of cognitive effort. Therefore, one should not assume that errors in reasoning occur due to a lack of cognitive effort, or an effortless process. Importantly, one cannot use the responses or performance on items from the heuristics and bias literature as a measure of cognitive effort.

[3] According to the dual-process explanation proposed by Kelemen et al. (2013) there could have been conflicting intuitions. This is further discussed in section 4.3 Comparing decision-making frameworks.

4.2 Comparing tools measuring cognitive effort

As reviewed in “the effort paradox” there is a need to compare and contrast tools measuring cognitive effort (Thomson & Oppenheimer, 2022). In Paper 1, we found that performance on three behavioral measures of cognitive effort, rational reasoning battery, DST and COG-ED, were not related to each other. However, we found that both COG-ED and rational reasoning battery were related to both working memory capacity, and a tendency to enjoy engaging cognitively demanding activity measured with the self-report scale, NFC. Additionally, the subjective mental demand experienced on each task was assessed on a visual analog scale with the N-TLX. In Paper 2 and Paper 3, we measured pupil dilation during task performance on two reasoning tasks as a physiological measure of cognitive effort. Finding more cognitive effort associated with errors in reasoning on a teleological reasoning task. On the contrary, more cognitive effort was related to avoiding base-rate neglect and avoiding reasoning errors on the base-rate task.

4.2.1 The need for cognition scale

The need for cognition scale (NFC) has been a widely applied measure of cognitive effort (Cacioppo et al., 1996). After more than 40 years since its invention it is still considered a globally valid trait measure of a thinking disposition to enjoy challenging mental activity (Cacioppo & Petty, 1982; Hussey & Hughes, 2020). As reviewed by Hussey & Hughes (Hussey & Hughes, 2020) the measure has good internal consistency, test-retest reliability, factor structure and measurement invariance. In Paper 1, I replicated that individuals higher in NFC show less effort discounting of monetary reward, i.e., less effort cost, in the COG-ED (Westbrook et al., 2013), replicating Westbrook et al. (Westbrook et al., 2013). Notably, the relationship between these tasks may be attenuated as individuals high in NFC have shown to be less responsive to reward incentives (Sandra & Otto, 2018). Furthermore, Paper 1 also found that individuals higher in NFC also perform better on rational reasoning tasks. This is consistent with multiple studies investigating the relationship between thinking disposition and performance on heuristics and bias tasks (Frederick, 2005; Thomson & Oppenheimer, 2016; Toplak et al., 2014; West et al., 2008). The NFC scale is an explicit cognitive effort measure as it directly asks about preferences for mental work, and enjoyment of cognitive effort (Fleischhauer et al., 2013; Strobel et al., 2015). The scale is likely measuring intrinsic motivation for cognitive effort, rather than extrinsic motivation as questions pertain to enjoyment and preferences for cognitive effort, or avoidance of cognitive work (reverse scored), without any question being related to extrinsic rewards or obligations. However, the scale seems to favor a preference for complex and challenging mental work rather than all cognitive work, e.g. “I would prefer complex to simple problems”. Indeed, the NFC has been associated with complex problem solving (Rudolph et al., 2018). This might partly explain why we found that the scale is not related to DST, as even the high demand deck involves quite simple problems that do not challenge most individuals intellectually. However, the NFC scale does include multiple questions related to avoidance of demanding cognitive work such as “I would rather do something that requires little thought than something that is sure to challenge my thinking abilities (reverse scored). Additionally, both measures gauge intrinsic cognitive effort. Thus, as the NFC is well established and did correlate with the other cognitive effort measures, the lack of a relationship between the two measures does bring into question the validity of the DST as a cognitive effort measure. The lack of a relationship between NFC and DST is consistent with Strobel et al. (Strobel et al., 2020). Notably, the NFC has also been applied together with pupillometry, finding higher NFC being negatively related to baseline (tonic) pupil diameter (da Silva Castanheira et al., 2021).

The need for cognition scale differs from the other effort measures reported in Paper 1 on multiple dimensions. Unlike the other effort measures used, NFC puts very little strain on the cognitive systems involved in cognitive effort such as executive function, working memory and attention. Although, NFC has been shown to be related to general intelligence, but not working memory nor executive functions (Gärtner et al., 2021; Hill et al., 2013). However, the measure is dependent on meta-cognitive awareness and self-awareness, which is a potential bias as introspective awareness may be limited (Nisbett & Wilson, 1977). Further, unlike the other behavioral measures of cognitive effort in Paper 1 there is no performance being measured with the NFC scale, thus the only measure is self-reported. Furthermore, NFC is considered a relatively stable trait (Cacioppo et al., 1996; Hussey & Hughes, 2020; Soubelet & Salthouse, 2017) whereas it is not known to what degree the other measures in our study change over time, but likely they are (more) state dependent. Notably, a state measure of motivation for cognitive effort has been developed for researchers interested in assessing state rather than trait NFC (Blaise et al., 2021).

In summary, the NFC is a reliable, valid, explicit, self-reported, trait measure of cognitive effort, which captures the intrinsic motivation to engage in and enjoy cognitive effort. Researchers should be mindful of what aspect of cognitive effort they are interested in measuring, in addition to general biases related to self-report questionnaires when applying NFC.

4.2.2 Rational reasoning battery

The rational reasoning tasks used in Paper 1 consisted of a mixture of problems selected from the heuristics and bias literature. The measure is an implicit performance measure of intrinsic motivation to expend cognitive effort. The measure is supposed to be an indicator of reflective capacity (Stanovich, 2009b), and be dependent upon working memory for performance (J. St. B. T. Evans & Stanovich, 2013; Kahneman, 2011; Pennycook et al., 2015a). However, in Paper 1 performance on the rational reasoning battery were not related to any other behavioral measure of cognitive effort. To the best of my knowledge Paper 1 is the first study investigating the relationship between rational reasoning tasks and the behavioral task paradigms, DST and COG-ED. The finding that these tasks measuring cognitive effort are not related is novel and has important implications as it has long been assumed that performance on these tasks were associated with cognitive effort (J. St. B. T. Evans, 2008; J. St. B. T. Evans & Stanovich, 2013; Frederick, 2005; Kahneman, 2011; Kahneman & Frederick, 2002; Shah & Oppenheimer, 2008; West et al., 2008). Importantly, subjective cognitive effort as rated with N-TLX showed that high performance on the rational reasoning battery was associated with less subjective effort, whereas lower performance was associated with higher subjective mental demand. This finding should strongly discourage researchers from using these tasks as a measure of invested cognitive effort. The finding of high performers reporting less cognitive effort is consistent with the smart intuitor dual-process account, suggesting that high cognitive ability individuals answer correctly without deliberation (Bago & De Neys, 2019; Raoelison et al., 2020; Raoelison & De Neys, 2019).

The relationship between rational reasoning and cognitive effort was assessed by measuring pupil dilation during task performance on two reasoning tasks in Paper 2 and Paper 3. In Paper 2 we found that more not less cognitive effort, as assessed by pupil dilation, was related to errors on a teleological reasoning task. This finding is consistent with Eldar et al. (Eldar et al., 2021) who also found errors on reasoning tasks being related to larger not smaller pupil dilations. Notably, Eldar et al. (Eldar et al., 2021) did not find a significant relationship between pupil dilation and performance on all tasks measured, but no task showed more effort being related to higher performance. Thus, while measuring concurrent cognitive effort during task performance these studies suggests that less effort, not more is related to higher performance on rational reasoning tasks. In contrast to this, in Paper 3 we did find that errors on a base-rate task was indeed related to smaller pupil dilations, i.e., less cognitive effort. Additionally, there was evidence that some pieces of information were neglected by a group of participants. Thus, for some rational reasoning tasks more effort could be related to higher performance. However, this thesis suggests that researchers should not assume that errors in reasoning occur due to a lack of cognitive effort. Rather, as shown in Paper 2, Paper 3 and by Eldar et al. (Eldar et al., 2021), individual tasks can be investigated by using psychophysiological measures such as pupil dilation and eye-tracking to elucidate the process underlying errors in reasoning.

Rational reasoning tasks have been proposed to measure both cognitive ability and a disposition (Frederick, 2005; Stanovich, 2009b, 2009a). Previous work has indeed found that rational reasoning tasks are related to thinking biases (Thomson & Oppenheimer, 2016; Toplak et al., 2014; West et al., 2008). Consistent with the literature, in Paper 1 we did find a relationship between rational reasoning battery and NFC. This association was also present when controlling for working memory capacity, measured with performance (discriminability) on the n-back task. Indeed, cognitive ability is an important factor that needs to be addressed when discussing the relationship between rational reasoning and NFC, or thinking disposition. As mentioned above, NFC has been found to be related to general intelligence, but not working memory (Hill et al., 2013), thus we cannot know if the relationship between NFC and rational reasoning battery found in Paper 1 would still be present if controlling for general intelligence instead of working memory. Notably, the CRT is a substantial portion of the rational reasoning battery applied in Paper 1 (up to 50%). A meta-analysis by Otero et al. (Otero et al., 2022) found the CRT could be explained by general intelligence and numerical ability, and no separate factor of cognitive reflection could be identified. However, the rational reasoning battery from Paper 1 consist of more than the CRT. Thinking dispositions have been found to correlate with such rational reasoning task batteries even when controlling for cognitive ability, by both Toplak et al. (Toplak et al., 2014) and West et al. (West et al., 2008), although the analyses can be criticized for not residualizing the variance (Croon, 2002; Hayes & Usami, 2020; Otero et al., 2022). Thus, rational reasoning tasks are related to cognitive ability. However, it is unclear to what degree these rational reasoning tasks measure something more than cognitive ability, they may also measure thinking dispositions. Further research is needed to disentangle what these rational reasoning tasks are measuring, and critically, there is a need to control for cognitive ability while doing so.

Importantly, this thesis suggests that rational reasoning tasks should not be applied as a measure of cognitive effort with the current state of knowledge. However, future research might be able to identify a set of tasks which do depend on cognitive effort by applying methods such as psychophysiological measures and computational modelling, similar to Paper 2 and Paper 3, as well as applied by Eldar et al. (Eldar et al., 2021) and Lin et al. (Lin et al., 2023)(2023).

4.2.3 Cognitive effort discounting paradigm

The cognitive effort discounting paradigm is an explicit measure of cognitive effort for external rewards. By applying a well-established working memory measure, the n-back task, cognitive demand can be manipulated by increasing processing quantity in the task (Owen et al., 2005; Westbrook et al., 2013). Consistent with previous research, Paper 1 replicated that cognitive effort discounting of monetary rewards, i.e., cognitive effort, is related to NFC and working memory capacity (A.-W. Kramer et al., 2021; Westbrook et al., 2013). As mentioned in the previous section, COG-ED was not related to rational reasoning. Additionally, Paper 1 found no relationship between DST and COG-ED. This was surprising as both COGED- and DST are commonly used behavioral measures of cognitive effort (Culbreth et al., 2016; J. M. Gold et al., 2015; Kool et al., 2010; A.-W. Kramer et al., 2021; Nagase et al., 2018; Sayalı & Badre, 2021; Strobel et al., 2020; Vogel et al., 2020; Westbrook et al., 2013, 2019, 2020; Zerna et al., 2023). The lack of a relationship between COG-ED and DST might be due to COG-ED being an explicit measure relying on external motivation, whereas DST is an implicit measure of internal motivation to avoid cognitive demand. However, the lack of a relationship between COG-ED, and rational reasoning battery and DST, might also be due to limitations of DST and rational reasoning battery as measures of cognitive effort (see the respective sections in 4.2 Comparing tools measuring cognitive effort). The COG-ED paradigm has successfully been applied with physiological measures such as fMRI, PET, and eye-tracking, and has helped elucidate the neural underpinnings of cognitive effort (Westbrook et al., 2019, 2020). The fact that the measure also provides an indicator of cognitive ability (n-back performance) is a strength of COG-ED. This allows for adjusting required performance to participants ability. As well as providing a measure to control for cognitive ability, which is convenient as cognitive ability is a confound in most cognitive effort research. Additionally, the COG-ED paradigm allows for varying load levels by adjusting processing quantity. Like most studies using the COG-ED paradigm, Paper 1 applied 1-back as a comparison task. However, 1 back may not be the optimal comparison task as some individuals might find the task boring and prefer some challenge. Indeed, Zerna et al. (Zerna et al., 2023) found that a significant portion (34.5%) of participants preferred higher load levels (although the majority preferred 1-back) and created a comparison paradigm which can account for differences in preferred task load. Additionally, the study found that discounting is not linear between load levels but is best described by a declining logistic curve. Discounting is steepest between 2-back and 3-back, and less steep between 1-back and 2-back, and 3-back and 4-back. Thus, a version applying 2-back as the comparison task for 3-back and 4-back may work as a faster COG-ED. Indeed, COG-ED was rated as the most subjectively effortful task paradigm in Paper 1. Furthermore, Zerna et al. (Zerna et al., 2023) showed that participants high in NFC perceived the highest levels of the n-back task as less aversive and less effortful. Thus, researchers need to be mindful when applying COG-ED as cognitive ability, reward sensitivity and individual differences in NFC may influence results (Sandra & Otto, 2018; Westbrook et al., 2013; Zerna et al., 2023). However, COG-ED has proved to be a reliable measure of cognitive effort. Future research should aim to evaluate different COG-ED task structures and paradigms to make effort assessments accurate and convenient.

4.2.4 Demand selection task

The demand selection task is an implicit behavioral measure of internal motivation to avoid cognitive effort (Kool et al., 2010). The results from Paper 1 showed an overall demand avoidance, although the effect was small. However, DST was not related to COG-ED nor rational reasoning battery in Paper 1. Perhaps most surprising was that DST was not related to NFC, despite both being intrinsic motivation measures of cognitive effort. Furthermore, the test-retest reliability of DST was poor. However, these results are consistent with Strobel et al. (Strobel et al., 2020) who found no relationship between NFC and DST, and questionable test-retest reliability of DST. The results from Paper 1 and previous studies raises concerns regarding this task paradigm, namely that choices on the DST may be influenced by color- and side preferences, demand avoidance is not always present, and demand avoidance is influenced by detecting the demand manipulation (J. M. Gold et al., 2015; Kool et al., 2010; Tran et al., 2022) Although, the DST produces a general demand avoidance, the version applied in Paper 1 may not be well-suited as an individual difference measure at the current state of knowledge. As the DST showed no relation to other cognitive effort measures it is hard to say what the task is measuring. However, there are multiple alternative administrations for this task paradigm which might make this task paradigm more reliable. Such as, changing the effort level by altering the frequency of rule changes between rounds, using forced trials, and applying rewards (Kool et al., 2010; Reddy et al., 2015; Sayalı & Badre, 2019). However, to date little is known about how the alternate versions of the task relate to cognitive effort and alternative cognitive effort measures. Thus, research is needed to validate versions of this task paradigm. Importantly, researchers should not assume similarities between this task paradigm and other cognitive effort measures.

4.2.5 NASA task load index

The NASA task load index has been widely used as a measure of workload and subjective cognitive effort for more than 30 years (Hart & Staveland, 1988). It has proven to be a valid and reliable measure (Braarud, 2021; Devos et al., 2020; Hart, 2006; Tubbs-Cooley et al., 2018). The measure can be applied both during and immediately after task performance. The N-TLX has been used to validate the COG-ED by providing subjective effort ratings for each load level of the n-back task (Westbrook et al., 2013). In Paper 1, the tool provided valuable information about the subjective cognitive effort associated with each task paradigm in their entirety. The measure also provided valuable information about the inverse relationship between rational reasoning performance and experienced mental effort. Researchers applying the N-TLX should be mindful that the measure is of subjective cognitive effort, which may differ from physiological or objective measures of cognitive effort (Kreis et al., 2020), and introspective awareness have limitations (Nisbett & Wilson, 1977).

4.2.6 Pupillometry

Pupil diameter can be applied as a physiological measure of cognitive effort (Hess & Polt, 1964; Kahneman & Beatty, 1966; van der Wel & van Steenbergen, 2018). In Study 2 larger pupil dilations were associated with errors on a teleological reasoning task, providing evidence that more not less cognitive effort was associated with errors in reasoning on this task. Additionally, larger baseline pupil size was related to better performance. This may indicate that sufficient arousal or intermediate pupil size and tonic LC activity may be optimal for task performance, consistent with the adaptive gain theory (Aston-Jones & Cohen, 2005). In paper 3, we found limited evidence that larger pupil size was related to conflict detection and cognitive decoupling as proposed by Pennycook et al. (Pennycook et al., 2015b), however this depended on the analyses used. There was limited evidence that more cognitive effort was related to higher performance on a base-rate task. Additionally, there was evidence suggesting that conflicting information elicited more cognitive effort compared to congruent information. Regarding baseline pupil, in Paper 3 smaller baseline pupil size was related to higher performance on the base-rate task. This could be explained as larger baseline pupil size indicating high tonic LC activity and "exploration" or distractibility according to the adaptive gain theory (Aston-Jones & Cohen, 2005), or mind-wandering (Mittner et al., 2016). Thus, the results from Paper 2 and Paper 3 are directly in contrast to each other regarding pupil dynamics and performance.

According to the adaptive gain-theory there is an inverted-U shaped relationship between arousal and performance. If participants baseline pupil size, arousal and LC activity were on opposite sides of the inverted U-shape in paper 2 and 3, we would expect on the one hand larger baseline pupil size being related to higher performance (Paper 2), and on the other hand smaller baseline pupil size being related to higher performance (Paper 3). There are important differences from an attentional point of view. Firstly. in the base-rate task half of the participants had already performed a similar version of the task. Secondly, the duration of the task and each trial was longer compared to the short statements presented in the teleological reasoning task. Thirdly, participants did not receive feedback in the base-rate task. Thus, it is conceivable that participants were more likely mind-wandering or distracted during the base-rate task. This highlights a challenge with pupillometry as a measure of cognitive effort.

Pupil size fluctuates spontaneously through time likely as a result of arousal and LC activity, but also other brain areas and neuromodulators such as acetylcholine influence pupil size (Aston-Jones & Cohen, 2005; Beatty & Lucero-Wagoner, 2000; Mathot, 2018; Reimer et al., 2016). Thus, changes in pupil size are almost impossible to link with certainty to one specific mental process. Additionally, as noted above, intermediate baseline pupil size is optimal for on-task performance and both too large and too small pupil size can be detrimental for performance (Aston-Jones & Cohen, 2005). Further complicating pupillometry as a measure of cognitive effort, is the interaction between baseline pupil size and pupil dilations. It can be inverse where larger pupil dilations and high tonic LC activity, is associated with smaller pupil dilations (Eldar et al., 2013; Gilzenrat et al., 2010). Alternatively, the relationship between baseline pupil size and pupil dilations also shows an inverted-U shape if investigating a larger specter of possible pupil sizes, as very small pupils are also associated with smaller dilations (Mridha et al., 2021).

That larger pupil dilations were associated with higher performance in the base-rate task but not the teleological reasoning task can be explained as higher requirements on executive function and cognitive control. Performance in the base-rate task required integration of multiple pieces of information presented sequentially for successful performance, leading to more information being held and updated in working memory over time, requiring more cognitive control and executive functions (Friedman & Miyake, 2017; van der Wel & van Steenbergen, 2018). The teleological reasoning task on the other hand, presumably did not differ in executive control required for correct and error responses. Although this was proposed by Kelemen et al. (Kelemen et al., 2013), the evidence suggested that fast and effortless processes were associated with correct responses on the teleological reasoning task.

Pupillometry has proven to be a valid physiological measure of cognitive effort (van der Wel & van Steenbergen, 2018). However, the pupil does not only measure cognitive effort. It is influenced by virtually any mental processing, in addition to light, near fixation and orienting (Beatty & Lucero-Wagoner, 2000; Mathot, 2018). Furthermore, spontaneous fluctuations in arousal influence pupil size and performance (Aston-Jones & Cohen, 2005; Chiew & Braver, 2013; Gilzenrat et al., 2010; Mathot, 2018; Reimer et al., 2016)). Lastly, fluctuations in baseline pupil size influence the size of event related pupil dilations (Eldar et al., 2013; Mridha et al., 2021). Thus, pupillometry and pupil dilations can be used as a physiological measure of cognitive effort. However, researchers need to be mindful when designing experiments by considering lighting, time on task, pace, timing of measurement, appropriate stimuli and control conditions, see Mathôt & Vilotijević (Mathôt & Vilotijević, 2023) for general guidelines. Importantly, I advise researchers to go a step beyond baseline correction as simply removing noise and consider the influence of baseline pupil size and arousal when measuring cognitive effort as event-related pupil dilations.

The application of pupillometry to investigate reasoning tasks has elucidated the role of cognitive effort in two reasoning tasks in this thesis and has provided valuable evidence for comparing decision-making frameworks. Except for Eldar et al. (Eldar et al., 2021), this methodology is novel and provides a path for research to further elucidate the role of cognitive effort in reasoning tasks and errors in reasoning.

4.3 Comparing decision-making frameworks

The study of decision-making and errors in reasoning has been influential in popular media, research, and has informed behavioral interventions and public policy (Buttenheim et al., 2023; De Neys, 2018; J. St. B. T. Evans & Stanovich, 2013; Kahneman, 2011; Kim et al., 2006; United Nations, 2021). Several decision-making paradigms have been proposed to explain errors in reasoning and deviations from rational behavior. However, these decision-making frameworks make opposing predictions regarding some decision biases and errors in reasoning. In this thesis the classical dual-process default-interventionist account, along with theoretical advancements such as the smart intuitor and the three-stage model of analytic engagement has been investigated alongside the extensive integration account of bias in reasoning, building on sequential sampling models of decision-making.

4.3.1 Dual-process models

A simplistic explanation of reasoning errors according to the default-interventionist dual-process account is that errors in reasoning occur through fast and effortless processes (or is at least more likely). Deliberation, which requires time and working memory resources, is associated with correct responses and fewer errors in reasoning. Thus, correct responses should be slower and require more effort, whereas errors in reasoning should be fast and require less effort. The smart intuitor account holds the same predictions, however, highlighting that correct responses can also be fast and effortless.

In Paper 1 we found that low performance on a rational reasoning battery, i.e., errors in reasoning, were associated with more subjective cognitive effort as measured with the N-TLX. This finding is directly opposing predictions from the default-interventionist dual-process account. The CRT which is a substantial portion (50% in some studies) of the tasks included in the rational reasoning task battery is supposed to measure deliberation (or cognitive reflection) and overriding of intuitive errors (Frederick, 2005). However, evidence from more than 350 participants in Paper 1 directly oppose that the task is measuring this. Additionally, dual-process theories propose that performance on the remaining heuristics and bias tasks should also be associated with more cognitive effort or Type 2 processing (J. St. B. T. Evans, 2008; J. St. B. T. Evans & Stanovich, 2013). Thus, this provides substantial evidence directly opposing the default-interventionist dual-process account of errors in reasoning. The finding of subjective effort being negatively related to performance on rational reasoning tasks are however compatible with the smart intuitor account. If some individuals have high performance and intuitively come up with the correct answer and some don’t find the correct answer it is likely that the latter group will work to find the correct answer and thus will report higher effort, although they fail and answer incorrectly. However, this would also mean that most errors are not intuitive but occur after deliberation. Which then begs the questions of whether the account has any predictive value and whether two processes are required to explain the results?

Performance on rational reasoning tasks were not related to any other behavioral measure of cognitive effort in Paper 1. According to the default-interventionist account one would expect that cognitive effort avoidance, or higher cognitive effort costs should be associated with more errors in reasoning. However, with the smart intuitor account the relationship between effort and reasoning errors is less clear, and this result neither favors nor opposes the smart intuitor account.

Need for cognition was positively related to rational reasoning performance in Paper 1. This result is consistent with dual-process theories as it is proposed that high NFC individuals are more likely to engage in the demanding deliberate Type 2 processes required for correctly solving rational reasoning tasks (J. St. B. T. Evans & Stanovich, 2013; Stanovich, 2009a). According to Stanovich’s tri-partite model (Stanovich, 2009a) NFC would be categorized under the reflective mind and individual differences in thinking disposition. This is not the same as engaging in Type 2 processing (algorithmic mind), but this thinking disposition is associated with more frequently engaging in Type 2 processing (Stanovich, 2009a). However, it is not clear that the rational reasoning tasks do measure something more than intelligence (Otero et al., 2022), although studies have tried to partial out the influence of intelligence (J. St. B. T. Evans, 2008; Toplak et al., 2014). According to Stanovich (Stanovich, 2009b, 2016), rational reasoning tasks measure usual performance, whereas tests of intelligence measure maximum performance. Although, based on subjective reports of cognitive effort, more effort on these tasks do not lead to higher performance. Thus, it is still unclear what rational reasoning tasks are measuring and if the relationship with NFC could be due to differences in intelligence.

In Paper 2 I found that the common reasoning error of accepting false teleological explanations were associated with longer response times and larger pupil dilations i.e., more cognitive effort. This result is directly opposing dual-process theories, both the default-interventionist and the smart intuitor account. One could argue that the results are possible under the smart intuitor account. However, this account would lose predictive value if the effort and time premises can be completely reversed. The results from Paper 2 are also consistent with the findings of Eldar et al. (Eldar et al., 2021) who found that larger pupil dilations were associated with more bias on three framing tasks. Overall, the results from Paper 1 and Paper 2 are largely incompatible with dual-process theory and most results directly oppose predictions from dual-process theory.

The base-rate task on the other hand showed mixed evidence regarding predictions from dual-process theory. There was an overall effect of larger pupil dilations being associated with correct responses across conditions. However, this was not significant in the incongruent condition. Thus, indicating that there is an overall effect of effort on performance, but this effect was not specific for reasoning errors. According to dual-process theory, the incongruent condition is the only condition where larger pupil dilations should have been necessary, however larger pupil dilations should be related to higher performance across conditions also. Importantly, an effect of more effort being linked to correct responses in the incongruent condition was almost significant for the stereotype responders. Thus, the results are inconclusive and there is a need for a replication with higher statistical power to detect if smaller effects exist. Thus, the evidence here does not support dual-process theory, however it does not contradict dual-process theory either.

Regarding conflict detection, the results showed larger pupil dilations in the attribute time window for stereotype responders when they subsequently gave the base-rate congruent (correct) response, which is opposite to their usual response. This partially supports the role of cognitive effort in overriding errors in reasoning. Additionally, base-rate responders showed larger pupil dilations in the attribute time window in the incongruent condition compared to the congruent condition, supporting conflict detection (note they were also larger in the neutral condition). Thus, the results find evidence supporting the construct of conflict detection. However, the results should be considered preliminary and there is a need to replicate the experiment with increased statistical power.

A novel finding in Paper 3 was that there were two groups of responders on the base-rate task, stereotype responders and base-rate responders. Importantly, the results showed that in the incongruent condition these two groups are fastest when giving their majority response i.e., stereotype congruent responses for stereotype responders, compared to when giving their minority response. That is, base-rate responders were slower when giving the stereotype congruent response. This is contrary to predictions of dual-process theory which proposes that stereotype information should be processed fast and effortless, whereas integrating the base-rate information should require time and cognitive effort. However, intuitive processing of base-rates has been shown previously (Newman et al., 2017; Pennycook et al., 2014). Additionally, a bi-modal distribution of responses in the incongruent condition was also found in Pennycook & Thompson (Pennycook & Thompson, 2012), and a re-analysis might reveal a similar separation of two groups of responders. Notably, across participants the usual effect of stereotype congruent responses being fastest was present in the gaze version of the base-rate task, which is consistent with the literature (Pennycook et al., 2014, 2015b; Pennycook & Thompson, 2012). However, this effect was reversed in the pupillometry version of the base-rate task, where the order of presentation for the attribute and base-rate information is switched. This is contrary to dual-process theory which supposes that stereotype processing should always be faster than base-rate integration. Thus, a body of work may have incorrectly assumed that stereotype congruent responses result from fast effortless processing, when task structure, stimulus selection and response preference may have influenced the result. Note that dual-process models have received criticism for assuming intuitive processing based on response times before (Krajbich et al., 2015; Pennycook et al., 2016). Furthermore, it should be noted that some dual-process proponents argue that the speed of processing is a correlate of the two types of processing and cannot be taken as evidence of Type 1 or Type 2 processing (J. St. B. T. Evans & Stanovich, 2013). Although others argue for the importance of response time in dual-process research (Pennycook et al., 2016). Evidence for a fast use of base-rates, statistics, logic, and other information processing types believed to only arise through Type 2 processing have accumulated in the last decade (Bago & De Neys, 2017; Newman et al., 2017; Pennycook et al., 2014; Raoelison et al., 2020; Thompson et al., 2018).

The results regarding the two groups, quite convincingly show that one group of responders do seem to neglect the base-rate information. The stereotype responders are insensitive to the base-rate information, and it does not affect response times, gaze, nor pupil dilations. These responders give the stereotype congruent response in the incongruent condition on the majority of trials. Thus, supporting that neglect of the base-rate information is a significant source of bias on this task. This is predicted by dual-process theory and could be considered evidence in support of the dual-process theory. However, what does it mean to say that this is a Type 1 process, and does it add any explanatory value?

Within the default-interventionist account, if a central distinction between the two types of reasoning is the mode of operation and the types of information and computations that can be performed by the two processes. E.g., Type 1 processing operates by association (or conditional and operant leaning principles), and form stereotypes quickly based on past experience, but cannot operate on formal logic, probabilities or mathematical principles. It adds explanatory value and is of interest that participants neglect some information (base-rates), whereas other information (forming stereotypes) is utilized (although, an autonomous process has not been proven), and may require less effort (although the results are mixed regarding this in Paper 3).

In the smart intuitor account, stereotype formation, logic and probabilities can be intuitive (Newman et al., 2017; Pennycook & Thompson, 2012; Raoelison et al., 2020; Thompson et al., 2018). Additionally, Type 2 processing can be belief based, improperly use statistics, lead to incorrect responses, and rarely lead to an improvement in responses (Ferreira et al., 2022; Newman et al., 2017; Pennycook & Thompson, 2012; Raoelison et al., 2020; Thompson et al., 2018). If this is the case, then it is unclear whether it adds any value to say that some process was Type 1 or Type 2. Additionally, there seems to be a tendency towards looking at the two types of processing as acting in parallel (Trippas et al., 2017; Trippas & Handley, 2018), or even opening up for the possibility of a continuum model (Kruglanski & Gigerenzer, 2011; Newman et al., 2017; Osman, 2004; Raoelison et al., 2020; Thompson & Newman, 2020). According to Stanovich and Evans (J. St. B. T. Evans & Stanovich, 2013), the defining feature of Type 1 processing is that it is autonomous, whereas Type 2 processing relies on working memory (and speed is just a correlate of the two). However, the tasks usually applied to investigate dual-process theory are rarely applied for such a distinction to be made (Thompson & Newman, 2020).

The results from this thesis overwhelmingly oppose predictions from the classical default-interventionist account, although some results are in line with the model. Within the smart intuitor account, many of the results could be accommodated under this model by post-hoc explanations of the results, but the model would lose predictive value and virtually become unfalsifiable. It is unclear if the distinction between the two processes is helpful under the smart intuitor account. If we accept that the output of a decision process cannot distinguish between the two processes through which it was made (J. St. B. T. Evans & Stanovich, 2013). And all types of information can be processed as both a Type 1 and Type 2 process depending on mindware or automatization (Stanovich, 2018). Is the most fruitful path forward for decision science to determine if a process was autonomous or relied on working memory? Or should efforts be focused on multiple aspects of the decision process? Such as what changes in the decision process when working memory load or cognitive effort is increased, or a conflict is detected? Does this change information sampling, breadth of attention, the influence of context, speed of accumulation or change of threshold? How does attention, memory and the internal representation of the problem evolve over time and interact with available information? Alternative approaches such as investigating reasoning strategies suggests that other factors than working memory may carry more explanatory value (de Chantal et al., 2020; Markovits et al., 2021; Thompson & Markovits, 2021).

Do the labels of Type 1 and Type 2 processing help advance research because they are useful constructs? I would argue these constructs have led to issues with communication as the constructs are too broad and general, and there is disagreement about their nature (J. St. B. T. Evans & Stanovich, 2013). Additionally, as these constructs are too broad and general researchers have struggled to be specific with their hypotheses and to specify the underlying mechanisms they are trying to investigate. I advise dual-process researchers to be more specific with their research questions and explicitly state the proposed underlying mechanism of investigation. Further, models should be tested against competing models, not a null model. Lastly, dual-process researchers could benefit from looking across research fields to expand their toolbox, notably psychophysiological measurements and computational modelling could aid researchers in specifying and testing competing models and the underlying neural structure.

It could be argued, perhaps rightfully so, that the portrayal of dual-process models has been simplified in this thesis. However, if the most basic assumptions of the models, that differences in cognitive effort and speed of responses are related to performance, do not hold in general, I argue that something is wrong with the model. This is important as dual-process models and decision-science have been highly influential, informing behavioral interventions and public policy (Buttenheim et al., 2023; Kahneman, 2011; Kim et al., 2006; United Nations, 2021). If the underlying theory is not correct it can lead to unintended consequences. For example, behavioral interventions aimed at increasing deliberation can have the intended effect, but not due to increased deliberation (Lin et al., 2023), or have the opposite effect (Van Gestel et al., 2021).

4.3.2 The three-stage model of analytic engagement

The three-stage model of analytic engagement proposes that multiple responses may be produced as a Type 1 process and that conflicts between these responses cause Type 2 processing. Further, the model separates between failure in conflict detection and cognitive decoupling as early and late sources of bias in reasoning (Pennycook et al., 2015b). The novel contributions of the model will be addressed in this section, for a discussion regarding Type 1 and Type 2 processing in general, see section 4.3.1 Dual-process models.

Conflict detection failure was a significant source of bias in Paper 3, where a large group of participants neglected the base-rate information. Thus, supporting the model assumption that conflict detection failure is a major source of bias in reasoning. Regarding conflict detection the evidence was mixed. There was difference in pupil dilation in the attribute time window between the congruent condition and the neutral condition but not between the congruent and incongruent condition as proposed by the model. Further, this difference was significant for the base-rate group but not the stereotype group. Thus, increased cognitive effort was found when the base-rates were non-informative. And this applied to the base-rate group but not the stereotype group. According to the model it is difficult to see how neutral base-rates should generate a response. Thus, the evidence seems to suggest that cognitive effort was engaged, or a "conflict was detected", not when there was conflicting responses. But perhaps when the base-rate group had to change strategy from base-rate information to the stereotype (class) information. This would be consistent with switching requiring cognitive effort (Friedman & Miyake, 2017; van der Wel & van Steenbergen, 2018), and also with LC activity being involved in reorienting or acting as a neural interrupt signal (Bouret & Sara, 2005; Dayan & Yu, 2006). However, these results do not support the hypothesis that conflicting responses causes Type 2 processing or analytic thinking (note that the model do not oppose other sources of analytic engagement). An addition to the model that would explain the result of the neutral condition being associated with larger pupil dilations is that an initial response or strategy could be evaluated as not leading to a correct answer, triggering a new process (not stating if this process would be Type 1 or Type 2 process, but rather asking if this process is similar or different, and how it is different, opening this up as a potential research avenue). Furthermore, it was found that stereotype responders did show larger pupil dilations, i.e., more cognitive effort, before giving the stereotype congruent response, which was opposite to their majority response. Indicating that the time window of conflict detection was important for correct responses and avoiding bias in reasoning. The same analysis barely missed significance when including the base-rate group. Thus, there is evidence that conflict detection and conflict detection failure may be important for avoiding bias on the base-rate task. However, we did not find evidence that conflicting intuitions were the source of conflict, triggering Type 2 processing. However, our data do not exclude that this could happen, and a larger powered study might detect a difference in pupil dilations in the attribute time window between the congruent and incongruent conditions as the direction of the effect was in the expected direction.

The evidence regarding cognitive decoupling in Paper 3 was mixed. When comparing pupil dilations before making a decision across conditions, base-rate responders showed larger pupil dilations in the incongruent condition compared to the congruent condition, whereas stereotype responders did not. However, the neutral condition also showed larger pupil dilations compared to the congruent condition for base-rate responders. Thus, for the base-rate responders it might be that integrating opposing evidence, or changing what type of information they were relying on might have required more effort. Additionally, there was an effect of larger pupil dilations before decisions being associated with correct responses across conditions. When looking solely at correct and error responses in the incongruent condition, there was no significant effect of pupil dilation before making a decision. However, the effect was in the expected direction and was nearly significant for the stereotype group. Thus, the evidence is ambiguous regarding cognitive decoupling. There might be smaller effects that Paper 3 were not able to detect. However, there may be a general effect of cognitive effort on performance. This is in accordance with previous work (Bonner & Sprinkle, 2002; van der Wel & van Steenbergen, 2018).

The three stage-model may be criticized as both Type 1 processes and Type 2 processes can both lead to errors and correct responses (see section 4.3.1 Dual-process models for a discussion on this point). The flexibility of the model is both a strength, as it can account for many different possibilities, and a weakness, as the flexibility makes it hard to make predictions. However, despite the flexibility of the model, there are a number of findings in Paper 3 the model cannot account for. Detecting a conflict is supposed to engage analytic thinking, which can be operationalized as increased decision threshold (Lin et al., 2023). However, modelling responses with a drift-diffusion model showed no difference in decision threshold. Rather the DDM suggested that response time differences across conditions is due to increased task difficulty, as indicated by lower drift-rate in the DDM. Further, the changes in response times due to the majority and minority responses are not easily explained by the model (although it is possible). Lastly, the changes due to task structure (gaze version and pupillometry version) in response times for correct and error responses in the incongruent condition is not possible to explain with the three-stage model of analytic engagement. However, the model adds valuable contributions by dissociating conflict detection failure and failure to come up with the correct response after having detected a conflict. Paper 3 overall supports the notion of dissociable sources of bias. Additionally, the proposal of multiple intuitions being generated internally, and this as a source of conflict that can influence subsequent processing is a valuable contribution, although not assessed here. Lastly, due to model flexibility the model could probably account for findings in Paper 1 and Paper 2. However, the criticism that the main predictions from dual-process models do not hold in general, i.e., more effort and longer deliberation time leads to better reasoning performance, is still valid and discredits the dual-process aspect of the model.

4.3.3 Sequential sampling models and the extensive integration account

Sequential sampling models are a class of computational decision-making models applied in different domains such as perceptual reasoning, memory, and value-based decision-making (Krajbich, 2019; Krajbich & Rangel, 2011; Ratcliff & McKoon, 2008). Additionally, they have been able to explain reasoning biases such as framing effects, loss aversion, preference reversals, and the similarity-, attraction-, and compromise effects (Busemeyer et al., 2006; J. G. Johnson & Busemeyer, 2005; Noguchi & Stewart, 2018; Trueblood et al., 2014; Tsetsos et al., 2012; Usher et al., 2013; Usher & McClelland, 2004). Notably, Eldar et al. (Eldar et al., 2021) noted that these models make the opposite assumption regarding pupil dilations and response time, compared to dual-process models. In paper 2, the theory of extensive integration as a source of reasoning bias was put to the test. Further, the drift-diffusion model was applied as a tool in Paper 2 and Paper 3 for investigation of latent parameters of the decision process.

The finding in Paper 1 that more subjective effort was associated with worse performance on reasoning tasks is in line with the extensive integration account, which predicts that more time and effort is associated with more bias. Note that Paper 1 did not set out to test sequential sampling models, nor the extensive integration account.

In Paper 2, we found that smaller pupil dilations and faster response times were associated with correct responses, i.e., less bias on a teleological reasoning task. Thus, our main outcome measures were in accordance with the predictions of the extensive integration account. Further corroborating the predictions from the extensive integration account, we found that larger baseline pupil size was related to correct responses. Indicating that smaller baseline pupil size, which was used as a proxy for low tonic LC activity and low neural gain, was associated with more bias in reasoning. Furthermore, results from analyzing trial-by trial variations in pupil size with the drift-diffusion model showed that smaller baseline pupil size and larger pupil dilations were associated with higher decision threshold and lower drift-rate, thus, requiring more evidence accumulation to make a decision. Conversely, larger baseline pupil size and smaller pupil dilations were associated with lower decision threshold and higher drift-rate, leading to less evidence accumulation, fewer errors in reasoning. Overall, the result from Paper 2 follows all predictions from the extensive integration account, providing convincing evidence in favor of this account of bias in reasoning. The application of modelling response times with a drift-diffusion model provided further evidence in favor of the extensive integration account and the link between pupil size, LC activity and neural gain (Aston-Jones & Cohen, 2005; Eldar, Cohen, et al., 2016; Eldar et al., 2013, 2021).

Paper 3 was not meant to test the extensive integration account of bias in reasoning. It is hard to make predictions for the extensive integration account for the base-rate task. On one hand, if there is a bias to attend the class/stereotype stimuli, this bias could be exacerbated by more time and extensive integration. On the other hand, lower neural gain could lead to slower, broader and more extensive integration, which includes integrating both types of information, which should result in less bias in reasoning. Thus, the task is not suited to make predictions regarding this account vs dual-process models, as post-hoc justifications for either result could be made.

The drift-diffusion model was applied as a tool in Paper 3 to assess if conflicting information did increase decision-threshold. An advantage of DDM's is that they decompose the entire response time-distributions into latent decision parameters, rather than relying on mean response times. The results from modelling the responses on the base-rate task with the DDM showed that there was no difference in threshold across conditions. Instead, the differences in response times across conditions could be explained by lower drift-rate, indicating higher task difficulty in the incongruent and neutral conditions compared to the congruent condition. This finding has important implications for dual-process research as mean differences in response times are often interpreted to indicate more deliberation or Type 2 processing (Bago & De Neys, 2017; De Neys, 2006; De Neys & Glumicic, 2008; J. St. B. T. Evans & Curtis-Holmes, 2005; Frederick, 2005; Kahneman, 2011; Pennycook & Thompson, 2012; Rubinstein, 2007). However, the fact that task difficulty may be manipulated through changing conditions is often overlooked. As highlighted in N.J. Evans & Wagenmakers (N. J. Evans & Wagenmakers, 2020) sequential sampling models (or evidence accumulation models) should be implemented as a default method for inference, rather than mean response times and accuracy rates.

Response bias was also found to differ between the gaze version and the pupillometry version of the base-rate tasks. This finding is hard to explain through dual-process models. However, if the decision process is viewed as a sequential sampling process this could be neatly explained. In the gaze version of the task participants receive first the class information, then the attribute. At this point a decision-process has likely started, where evidence is accumulated towards the class/group which is most congruent with the attribute information, i.e., the stereotype. Thus, when the response slide is presented together with the base-rate information, evidence has already started accumulating towards the group that is congruent with the attribute/stereotype, resulting in a starting point bias in the model towards the stereotype congruent option. Conversely, in the pupillometry version the task presentation order is switched. In the pupillometry version participants see the class/groups, then they are presented both the groups and the base-rate information. At this point evidence accumulation towards the higher base-rate group likely starts, as there is no stereotype information available. Then, the attribute information is presented. Lastly, a question regarding which group the person most likely belongs is presented. The results show that the starting point bias in this version is towards the base-rate congruent response. This could be due to the fact that this information is available earlier, and accumulation towards this option can start sooner. Alternatively, the change of modality can matter. The participants see both the class and base-rate information on the same slide, whereas they have to remember the two groups when the attribute is being presented. The exact mechanism is not known. However, sequential sampling models can provide a simple explanation for why the starting point bias change when task structure is alternated. Additionally, the change in response times in the incongruent condition between correct and error responses for base-rate responders and stereotype responders can be explained through the same mechanism as laid out in Krajbich et al.'s (Krajbich et al., 2015) criticisms of dual-process research.

It is beyond the scope of this thesis (and Paper 3) to provide formal models of the base-rate task with sequential sampling models. However, it might be fruitful to ask how visual attention influence the decision process, perhaps with an adopted version of the attentional drift-diffusion model (Krajbich & Rangel, 2011; Yang & Krajbich, 2023). Or one could ask why the stereotype/class information seem to draw more attention, is it due to being more uncertain, and it is optimal to pay more attention to options that are high in value and uncertain (Callaway et al., 2021).

Outside of Paper 2, there was no strong tests of sequential sampling models or the extensive integration account of bias in reasoning in this thesis. Therefore, sequential sampling models have not been tested or evaluated to the same degree as dual-process models have in this thesis. To my knowledge there is no strong evidence against sequential sampling models in the papers in this thesis. However, sequential sampling models can be criticized for being to restrictive and not applicable to complex problems such as those studied in the heuristics and bias literature. It is true that the models are limited in the timeframes they are applicable for and depend on sufficient response data. However, Paper 2 and Paper 3 provide evidence that these models can be applied to reasoning problems by adapting the presentation format of the problems. Additionally, the models can be criticized for being too general and do not provide in-depth task specific explanations of cognitive phenomena. However, task-specific extensions of the models can overcome this issue.

4.4 The role of cognitive effort in decision-making

Decision-making paradigms have been employed to study cognitive effort costs and cognitive demand avoidance (Kool et al., 2010; Westbrook et al., 2013). In general, cognitive effort is expected to have a beneficial effect on task performance and decision making (Bonner & Sprinkle, 2002; G. R. Hockey, 1997; Shenhav et al., 2013; Shenhav, Musslick, et al., 2017; van der Wel & van Steenbergen, 2018). Additionally, it has been proposed that a lack of cognitive effort is associated with errors in reasoning (J. St. B. T. Evans, 2008; J. St. B. T. Evans & Stanovich, 2013; Frederick, 2005; Kahneman, 2011; Kahneman & Frederick, 2002).

Paper 1 assessed two common behavioral measures for investigating cognitive effort through decision-making. The COG-ED investigates cognitive effort through decisions to expend more cognitive effort for a larger reward or less effort for a smaller reward. The task paradigm was reported as being subjectively effortful and provided reliable individual difference measures of cognitive effort discounting of monetary rewards. The task was related to working memory performance and NFC, thus replicating previous work (Westbrook et al., 2013). The task was assessed as being a valid paradigm for studying cognitive effort through decision-making. The DST studies cognitive effort implicitly through studying a series of behavioral choices between two decks, which unbeknownst to participants, differ in cognitive effort. The DST did show cognitive effort avoidance as there was a small overall tendency to avoid the high cognitive demand deck, replicating Kool et al. (Kool et al., 2010). However, the task paradigm was not related to any other cognitive effort measures in Paper 1. Thus, the task paradigm shows cognitive demand avoidance, but the task version applied in Paper 1 may not be suitable as a paradigm to study individual differences in cognitive effort through decision-making. Overall, decision-making paradigms can be a valuable tool for studying cognitive effort, however researchers need to be mindful that cognitive effort is not a unitary construct and different task paradigms measure different aspects of cognitive effort and have different strengths and weaknesses (see section 4.2 Comparing tools measuring cognitive effort).

In paper 3, we found an overall effect of larger pupil dilations and more cognitive effort being related to better performance across conditions. This is in line with previous work showing a general effect of higher cognitive effort being linked to better performance (Bonner & Sprinkle, 2002; G. R. Hockey, 1997; Shenhav et al., 2013; Shenhav, Musslick, et al., 2017; van der Wel & van Steenbergen, 2018). However, researchers investigating cognitive effort and performance need to be mindful of the influence of arousal which can be detrimental to task performance at both too high and too low levels (Aston-Jones & Cohen, 2005). Further, arousal will influence neural gain and pupil size, which can impact performance (Eldar et al., 2013; Reimer et al., 2016). Additionally, different levels of arousal may benefit different cognitive operations (Berridge & Spencer, 2016; Spencer & Berridge, 2019).

Dual-process theories generally assume that more cognitive effort should be associated with better performance on reasoning tasks, and a lack of cognitive effort is associated with more bias and errors in reasoning (J. St. B. T. Evans & Stanovich, 2013; Kahneman, 2011; Kahneman & Frederick, 2002). Contrary to this prediction, we found that performance on rational reasoning tasks was negatively associated with subjective reports of cognitive effort demand on these tasks. Additionally, in Study 2 we found that larger pupil dilations indicating more cognitive effort was associate with more acceptance of false teleological explanations for natural phenomena, i.e., more reasoning bias. This is in line with a body of work contradicting the predictions of the default-interventionist dual-process account (Bago & De Neys, 2017; Eldar et al., 2021; Newman et al., 2017; Raoelison et al., 2020). However, in Paper 3 we found some evidence of more cognitive effort after detecting a conflict being related to overcoming reasoning errors. Thus, the evidence indicates that more cognitive effort could lead to overcoming some reasoning errors, however more effort could be related to more reasoning errors on other tasks. The evidence therefore quite clearly shows that errors in reasoning is not generally due to a lack of cognitive effort, and one cannot assume that errors in reasoning occur due to a lack of cognitive effort. There is therefore a need for future research to determine which errors in reasoning occur due to a lack of cognitive effort and disentangle cognitive effort from other factors leading to bias and errors in reasoning.

In summary, cognitive effort is generally linked to higher task performance in decision-making tasks (E. K. Miller & Cohen, 2001; Shenhav et al., 2013; Shenhav, Musslick, et al., 2017; van der Wel & van Steenbergen, 2018). However, there are exceptions to this, and the evidence clearly shows that it is not warranted to assume that errors in reasoning occur due to a lack of cognitive effort. Indeed, some errors in reasoning are associated with more, not less, cognitive effort (Eldar et al., 2021). This indicates that general assumptions of the default-interventionist dual-process model do not hold. Furthermore, studying cognitive effort through decision-making is a promising research avenue. However, there is a need to validate that the tasks are indeed measuring cognitive effort and specify the aspects of cognitive effort that is being measured, and further investigate how the cognitive effort measures are related to other measures of cognitive effort, cognitive ability and motivation.

4.5 Limitations and future directions

Limitations of the presented Papers have been noted in the preceding sections of the discussion, however there are some overall factors which should be highlighted in a more structured manner.

4.5.1 Task- selection and structure

Task selection and task structure are important factors in research design and have important implications for the outcome of empirical studies and the inferences that can be drawn from them.

Paper 1 highlights the importance of cross-validating research tasks measuring related constructs. Notably, task selection is an important factor to consider in Paper 1. Firstly, there are alternative tasks that were not included such as the cognitive effort expenditure for rewards task (Lopez-Gamundi & Wardle, 2018). This task was not included as we were not aware of this task when starting our data collection. Additionally, we did not include measures of physical effort and perceptual effort (Horan et al., 2015; Reddy et al., 2015). Secondly, the tasks included come in several variants and it is not known how different variants of the task paradigms would relate to each other. As an example, the COG-ED used explicit labeling of the N-back levels, rather than associating the levels of N with colors (Westbrook et al., 2013). This made the task choices more explicit regarding the demand manipulation. It is not known if, or how, the manipulation of associating levels of N with colors influence choices. There is a possibility that making the demand manipulation less obvious would make the task more similar to the DST, which might influence covariance between the tasks. Additionally, it is not clear how subjective indifference points in COG-ED are influenced by applying higher load levels such as N = 5 or N = 6, as was done in Study 1 and Study 2 in Paper 1. As the load discrepancy between tasks is larger it is probable that subjective indifference points may become higher. Further, it is possible that higher load levels are more influenced by error rates compared to lower levels of N (Zerna et al., 2023). Furthermore, the DST has been used in several variants (Kool et al., 2010; Reddy et al., 2015; Sayalı & Badre, 2019). And it is unknown if using an incentivized version of the DST, or by changing the demand manipulation, or adding more demand levels, the DST might become a better measure of individual differences in cognitive effort expenditure. This might be related to other cognitive effort measures. Furthermore, boredom might be a factor in these tasks.

Preferences for cognitive work vary between individuals and some individuals prefer higher levels of cognitive effort (Bustamante et al., 2023; Cacioppo & Petty, 1982; Zerna et al., 2023). This is an important point as the low demand tasks in Paper 1, COG-ED (N = 1) and DST (10 % task switch) had very low demand. Some participants may have found the low demand tasks boring and preferred the higher demand tasks. Zerna et al. (Zerna et al., 2023) proposes an approach where the preferred cognitive effort level is not assumed but rather assessed prior to monetary discounting choices in COG-ED. However, Zerna et al. (Zerna et al., 2023) found that most participants favored the lowest cognitive effort option (N = 1). Notably, the discounting was shallower between N = 1 vs. N = 2, compared to N = 2 vs. N = 3 choices. Thus, a simple approach to account for effects of boredom might be to have the N = 2 as the low demand option. Alternatively, assess the preferred load level as in Zerna et al. (Zerna et al., 2023). Furthermore, higher NFC may be associated with higher demand preference (Zerna et al., 2023), but lower reward sensitivity (Sandra & Otto, 2018). Thus, future research may therefore benefit from including measures of reward sensitivity when using incentivized cognitive effort paradigms.

The rational reasoning task battery consisted of a selection of problems taken from the heuristics and biases literature. However, this selection was based on previous studies applying these tasks and not a complete assessment of all possible tasks and a stringent selection based on formalized criteria. Thus, the selection of tasks was based mostly on previous application and could be argued was somewhat arbitrary. The validity of such an approach is therefore up for discussion. The application in Paper 1, aiming to assess behavioral reasoning tasks as a measure of cognitive effort against other measures of cognitive effort, I argue is valid. The tasks are indeed used as measures of analytic-, deliberate-, Type 2-, thinking/processing and are assumed to measure cognitive effort or a cognitive style of expending cognitive effort (J. St. B. T. Evans, 2008; J. St. B. T. Evans & Stanovich, 2013; Frederick, 2005; Kahneman, 2011; Kahneman & Frederick, 2002; Shah & Oppenheimer, 2008; Trippas et al., 2015; West et al., 2008). However, the results from Paper 1 suggests that these tasks should not be applied as such. Further criticism might be directed toward the application of N-TLX after completing entire cognitive effort task paradigms. One could argue that it would be more beneficial to measure subjective cognitive effort after individual N-back tasks as in Westbrook et al. (Westbrook et al., 2013), or after each individual reasoning problem in the rational reasoning task battery. Indeed, this would be a more nuanced approach with more data regarding cognitive effort experienced under different task loads and demands. However, this approach would require a substantial increase in N-TLX measurements for participants to fill out, which would require more time, more repetition and would probably have a negative influence on motivation to complete the experiments. Secondly, this would not allow for comparisons of how subjectively effortful the task paradigms are compared to each other. However, Paper 2 and Paper 3 shows that different reasoning biases may be differentially influenced by cognitive effort and further research is necessary to determine which tasks depend on cognitive effort and what the other contributing factors in reasoning bias is. Future research efforts could inform what tasks should be applied together and which should not. Importantly, the tasks selected and the task structure in Paper 2 and Paper 3 also influence the inferences that can be drawn in the individual papers and this thesis.

In Paper 2 the teleological reasoning task was adapted for concurrent pupil measurement. For this purpose, the task had to be adapted to include a time limit. This time limit was longer than speeded trials in Kelemen et al. (Kelemen et al., 2013), and the mean response times indicated that time was not an issue. Additionally, we found that errors were slow on the task, not fast, indicating time was not a major factor and errors were not impulsive. However, including a variant of the teleological reasoning task without a time limit would have allowed further disentangling of the effects of speed vs. accuracy, and bias as accumulating over time or occurring due to fast processing. Furthermore, the time-window for analyzing the decision was externally imposed and only peak dilation was utilized in further analyses. Applying a version without time restrictions and with analyses across the full timeseries of pupil dilation could reveal differences in the reasoning processing in terms of attention, effort, uncertainty and mind-wandering (Mittner et al., 2016; Unsworth & Robison, 2016, 2017; Urai et al., 2017). However, this would likely exclude drift-diffusion modelling of responses and make comparisons between the competing accounts of reasoning bias more complex. Another concern regarding task selection in Paper 2 is the inclusion of rational reasoning tasks as a measure of cognitive ability. In paper 1, this was used as a measure of cognitive effort. However, Paper 1 showed that this is not a valid measure of cognitive effort. The application of reasoning tasks as measures of cognitive ability can be justified as Otero et al. (Otero et al., 2022) found that the cognitive reflection test does indeed index cognitive ability and has been applied as a measure of cognitive ability in other studies (Raoelison et al., 2020). However, we do note that it is not entirely clear what this task battery of rational reasoning tasks measure, there is uncertainty involved when utilizing this measure, and caution should be made with interpretations.

In Paper 3 two variants of the base-rate task were applied. Both versions of the task did include many exposures to the same type of problem with conditions providing small differences in the presented task. It is not clear how multiple exposures to the same problem type influence responses. Previous work has shown that responses are influenced by which type of information is presented (or not presented), the order of the presented information and the extremity of the base-rates (Koehler, 1996; Pennycook et al., 2014, 2015b; Pennycook & Thompson, 2012). Notably, Paper 3 did not include a non-informative stereotype condition, which possibly could have influenced the stereotype responder’s strategy. Further, using moderate base-rates may have influenced both participant groups by base-rate responders becoming more likely to give the stereotype congruent response. Or stereotype responders becoming more aware of the base-rates and thus weight them more heavily. Perhaps it would not be possible to separate two groups of responders. Multiple exposures to the same problem may have resulted in participants choosing a strategy for solving the task. This is consistent with the results showing two types of responders. Recent research has highlighted the influence of reasoning strategy in rational reasoning problems (de Chantal et al., 2020; Thompson & Markovits, 2021). Future research could help disentangle how task structure and presentation influence reasoning strategy and changes in reasoning strategy. However, task length is a limiting factor in task design as participants might get bored, which may influence performance, attention and pupil size (Aston-Jones & Cohen, 2005; Mittner et al., 2016; Unsworth & Robison, 2016). The finding of larger baseline pupil size being related to errors on the base-rate task in Paper 3, indicates that boredom, distractibility, or mind-wandering may have been a factor in Paper 3. Suggesting that there is a limit to the possible combinations that can be utilized in the base-rate task in single experiments. An additional factor in the design of the pupillometry version of the base-rate task was the change in modality during trials of the task, from information being presented visually for the two groups and the base-rate information, to auditorily for the attribute and question. This likely put a larger demand on working memory as more information had to be retained in working memory compared to the gaze version of the task. The reason for this change was to maintain illuminance and avoid light related changes in pupil size. However, increased working memory load may have increased cognitive effort expenditure, and the influence of motivation and cognitive ability.

4.5.2 Pupillometry

Pupil size was applied as a physiological measure of cognitive effort in Paper 2 and Paper 3. However, in Paper 2 it was also applied as a proxy for neural gain and activity of the LC-NE system. The pupil may be a “window to the preconscious” as noted by Laeng et al. (Laeng et al., 2012), providing valuable information about cognitive processing. However, the difficulty with using pupil dilation as a measure of cognitive processing is that anything that activates the mind or increases processing load also causes the pupil to dilate (Mathot, 2018). Thus, it is nearly impossible to link pupil dilation to a single cognitive action. In this thesis, it has been noted that pupil size may reflect cognitive effort, arousal, LC activity, neural gain, uncertainty, and mind-wandering. Notably, these concepts and neural structures are likely interrelated. Very low levels of arousal are likely associated with little cognitive effort, rising uncertainty may be accompanied by engagement of cognitive effort, and as mentioned, LC activity may influence neural gain and regulate task engagement. However, the issue of not knowing what one is measuring remains. Only briefly mention earlier is the fact that changes in pupil dilation are influenced by acetylcholine. Whereas LC phasic activity causes rapid changes in pupil size, longer lasting fluctuations in pupil size are influenced by cholinergic activity, such as those produced by locomotion (Reimer et al., 2016). Thus, in the case of the experiments in Paper 2 and Paper 3, where participants are seated with only small motor movements such as pressing a button, the systematic changes in pupil dilation are more likely related to LC activity. However, it cannot be excluded that cholinergic activity may have had a small influence on the results, or at least contributed to noise in the data. Further, fluctuations in arousal and LC activity may influence performance and task evoked pupil dilations (Aston-Jones & Cohen, 2005; Chiew & Braver, 2013; Eldar et al., 2013; Gilzenrat et al., 2010; Mathot, 2018; Mridha et al., 2021; Reimer et al., 2016). Thus, simple baseline corrections may hide information and influence results if not considered as a relevant or contributing factor. Designing experiments and task for concurrent measure of pupil dilation therefore requires careful consideration by researchers.

The timing of the pupil response is an important factor in pupillometry research. Ideally, there should be no external activity while pupil size is being measured (Mathôt & Vilotijević, 2023). The delay of the pupil response should be taken into consideration. In Paper 2 and Paper 3 the time window of pupil dilations before decisions (ending with the decision) may have been short, as the pupil response likely did not have time to fully develop. However, motor actions such as pressing a button also cause pupil dilation. Thus, the time window utilized was a compromise between capturing as much of the processing activity or intensity, i.e., cognitive effort leading up to the decision and avoiding dilation caused by motor-planning and motor action. An alternative approach could have been to instruct participants to delay responding by 2 seconds after making their decision. However, delaying the decision could be considered a factor influencing the decision (Chen & Krajbich, 2018; Martiny-Huenger et al., 2021; van de Ven et al., 2010) and would interfere with response time measures and drift-diffusion modelling. However, timing of the attribute time window in the base-rate task may have provided a good measure of mental activity related changes in pupil dilation, as participants waited for 2 seconds after the attribute presentation with no changes in the environment.

The stimuli should ideally be constant between conditions when designing pupillometry experiments. In the base-rate task, both the visual stimuli and the audio stimuli had minor differences in length (class and attribute), but these differences were small and not systematic. In the teleological reasoning task however, the differences in length of the statements presented varied with 1.4 seconds between the shortest and the longest audio clips (statements). The difference between listening for 2.3 second and 3.7 seconds most likely cause differences in pupil dilation and cognitive effort regardless of content. This is a confounding factor in the experiment that was not controlled for in the analyses. However, there are no known systematic differences in sentence length in the different statement categories. Another factor not yet discussed in the teleological reasoning task is that participants received feedback. Errors on the teleological reasoning task was accompanied by the expected increase in pupil dilations (Urai et al., 2017). It is possible that error trials on the task was accompanied by increased cognitive effort on the subsequent trial (Murphy et al., 2016). However, randomization of the items likely attenuates any systematic effect across trial types in pupil dilation and performance from error related engagement. Additionally, Paper 2 did not control for a learning effect due to feedback. However, a learning effect would not alter the predictions tested in Paper 2 and therefore not be consequential for the main results. Furthermore, it is possible that feedback in the teleological reasoning task may have had a positive effect on participants motivation (Burgers et al., 2015).

Recordings of pupil size are noisy even with high quality recordings, this is due to blinking, measurement errors and more (Mathot, 2018; Mathôt & Vilotijević, 2023). In both Paper 2 and Paper 3, there was loss of data due to low quality pupil data. The quality of the pupil recordings varied substantially between individuals, leading to exclusions of participants from analyses of pupil dilation. The sample size calculations had accounted for some loss of data due to pupil measurement inaccuracies. However, this did affect the statistical power of Paper 2, and in particular Paper 3. The findings from Paper 3 regarding pupil dilation should be considered preliminary as there were small effects in the expected direction that barely missed significance, and the level of significance likely depended on the analysis approach. To attain the most reliable data for publication, participants with too low quality pupil recordings were excluded. In addition, to remain transparent the supplementary analyses included additional analyses without exclusion of individual participants. However, the low quality of pupil recordings for some participants is a weakness of Paper 3. Additionally, the loss of data prevented exploration of interacting effects between individual difference measures such as cognitive ability and pupil measures. As mentioned previously, studies have shown an interaction between pupil dilation and cognitive ability (Granholm et al., 1996; Kreis et al., 2020; Poock, 1973).

4.5.3 Application of the drift-diffusion model

The drift-diffusion model was applied in Paper 2 and Paper 3 to investigate differences in responses across conditions. Additionally, in Paper 2 the pupil measures were included in the model as predictors of trial-by-trial variation in decision parameters. It has been recommended that the DDM should be applied to forced two-choice tasks with mean response times below 1.5 seconds (Ratcliff & McKoon, 2008). Before data collection it was not known if mean responses time on the teleological reasoning task would be below 1.5 seconds. However, Lerche & Voss (Lerche & Voss, 2019) have validated the DDM with tasks requiring longer response times. Mean response time in both the teleological reasoning task and the base-rate task was well below 1.5 seconds indicating no issues due to duration of responses in the two tasks. Additionally, it has been shown that the model parameters can be estimated with trial numbers as low as 50 – 200 with sufficient reliability (Lerche et al., 2017; Lerche & Voss, 2017, 2019). Thus, estimation of parameters should be reliable in Paper 2 (77 trials) and Paper 3 (80 trials). Notably, there was no convergence issues in estimating the parameters of the models and the model fit was good. The DDM has previously been applied to perceptual discrimination, lexical decision making, recognition memory, sentence comprehension, probabilistic- and reward- learning tasks, and decisions of cognitive effort and reward (Cavanagh et al., 2014; Germar et al., 2016; Ratcliff, 1978; Ratcliff & McKoon, 2008; Spaniol et al., 2008; Westbrook et al., 2020). The teleological reasoning task has similarities to sentence comprehension, lexical decision-making and recognition memory tasks. Decisions regarding a statement's veracity depends on knowledge about the world (recognition memory), detecting if information is not true or is flawed (similar to lexical decisions regarding detecting words vs. non-words), and inferring meaning from a statement (similar to sentence comprehension or understanding metaphors). Thus, there is precedence for applying DDM's to comparable tasks to the teleological reasoning task. Regarding the base-rate task there is no direct precedence for analyzing responses with a DDM. However, a comparable task is the effort decisions made in a COG-ED paradigm by Westbrook et al. (Westbrook et al., 2020). In this paradigm the participants made choices regarding which N-back and reward combination they wanted to perform, high effort and high reward vs. low effort and low reward. Choice options were vertically divided on the screen (left and right, similar to the base-rate task), and the information types (effort demands and rewards) associated with each option is horizontally divided, similar to the base-rate task. Thus, in both tasks the different information types favor opposing options, and one may be more heavily weighted than the other in the decision-making process. Furthermore, if the base-rate task had been divided into two sub tasks, making choices regarding if it is more likely that a person comes from a group of 3 vs. 997 people, it would be comparable to a probabilistic selection task (Cavanagh et al., 2014). Similarly, if the task was to state if an attribute (e.g., kind) was most likely describing one out of two groups (lawyers or nurses), the underlying process is comparable to a combination of recognition memory, perceptual discrimination, or value task (B. Forstmann et al., 2010; Krajbich et al., 2010; Krajbich & Rangel, 2011; Ratcliff, 1978). Thus, there are comparable task paradigms that have applied DDM's to analyze responses. However, it could be argued that the attentional drift diffusion model or a multi-attribute drift diffusion model might have been more appropriate for analyzing responses on the base-rate task (Fisher, 2021; Krajbich et al., 2010; Krajbich & Rangel, 2011). However, for a simple comparison of the main parameters of the model, I would argue that applying a standard DDM is sufficient.

4.5.4 General limitations of the thesis

This thesis has provided evidence that cognitive effort can be associated with more or less errors and bias in reasoning. As such it has provided evidence against the default-interventionist dual-process model. Further, tools measuring cognitive effort have been compared against each other. However, there are notable limitations to this thesis.

All cognitive effort measures applied in this thesis have their shortcomings and there is no objective benchmark to be compared against. Thus, uncertainty remains after comparing cognitive effort measures against each other. The uncertainty of what is being measured also applies to physiological measures of cognitive effort such as pupil dilation. Therefore, strengths and weaknesses of paradigms have been highlighted but there is no clear answer to how one should measure cognitive effort. As cognitive effort is a multi-faceted construct a single objective measure is likely not possible to find. As highlighted in the “effort paradox” (Thomson & Oppenheimer, 2022), the best current approach is to integrate research across fields to highlight different aspects of cognitive effort.

The interaction between cognitive ability and cognitive effort is an important factor in performance and for cognitive effort research. However, none of the papers in this thesis applied a valid measure of general intelligence. Rather, proxy measures such as working memory capacity were applied. Thus, there are questions regarding the interaction between cognitive ability and cognitive effort measures which remain unanswered.

The default-interventionist dual-process model was tested in several ways in this thesis finding strong evidence against this theory. However, sequential sampling models and the extensive integration account of bias in reasoning was not tested in the same manner. Apart from Paper 2, drift-diffusion models were only applied as a tool. Therefore, sequential sampling modes have not been properly tested in this thesis as a framework of decision-making. Thus, this thesis should not be taken as strong evidence in favor of these models.

The role of effort in decision-making and errors in reasoning still have many unanswered questions. This thesis showed that both more and less cognitive effort can be associated with errors in reasoning. However, the thesis does not provide an overarching theory for when effort will benefit decision-making and when it will be detrimental.

4.5.5 Future directions

The evidence from this thesis clearly opposes predictions from the default-interventionist dual-process model. This is in line with an accumulating body of research over the past decade. Notably, the thesis present evidence showing that one should not assume errors in reasoning generally occur due to a lack of cognitive effort. This presents an opportunity for future research to investigate reasoning errors separately, to investigate if cognitive effort is a contributing factor in making specific errors. Importantly, researchers should investigate alternative factors contributing to the specific reasoning errors apart from, or in addition to, cognitive effort. Furthermore, the labels of Type 1 and Type 2 reasoning may be too broad and lead to hypotheses and explanations that also suffer from being too broad and general. This can create an illusion of understanding a phenomenon by putting a label on it or presenting an “explanation”, without the underlying mechanism being explicit or tested. I advise researchers to question if their hypotheses could be more specific and consider if a single process could explain the phenomena at hand before applying the labels of Type 1 and Type 2 processing. Furthermore, researchers should aim to be specific in their hypotheses and test competing alternative explanations against each other, rather than relying on null hypothesis testing.

In this thesis, a combination of tools has been applied, such as pupillometry, eye-gaze and computational modelling. All these tools have previously been applied separately to investigate errors in reasoning. However, the combination of these tool, applied to reasoning problems is novel. Future, research on reasoning problems and cognitive effort could benefit from applying a variety of tools from related fields of research to advance the studies of decision-making, reasoning biases, heuristics, and cognitive effort.

This thesis has exposed weaknesses in several cognitive effort paradigms, and low correlations among cognitive effort measures. Thus, there is a need to create and validate new cognitive effort measures, and further evaluate and cross-validate alternative versions of existing task paradigms. Further, there is a need to communicate explicitly regarding the cognitive effort "construct". What part of cognitive effort is being measured, how is this (or is not) related to other cognitive effort measures and constructs? The framework outlined in the "effort paradox" could serve as a starting point (Thomson & Oppenheimer, 2022).

The involvement of the LC-NE system in cognitive effort and errors in reasoning suggests that a range of factors may influence decisions. Notably, the influence of arousal, sleep, and medications acting on the LC-NE system may have significant effects on decision-making and errors in reasoning. Future studies could manipulate levels of NE to provide further tests of the underlying theories regarding evidence accumulation, neural gain, and bias in decision-making. As the LC-NE system is involved in regulating arousal and attention further knowledge regarding the involvement of the LC-NE system in decision-making has widespread potential applications, e.g., air-traffic controllers, psychiatric disorders, medications, and more.

Lastly, the default-interventionist dual-process model has spread beyond research, influencing popular culture and behavioral interventions (Buttenheim et al., 2023; De Neys, 2018; Kahneman, 2011; Kim et al., 2006; Thaler & Sunstein, 2008; United Nations, 2021). If the underlying idea of reasoning errors resulting from a lack of cognitive effort is wrong this should be communicated to a wider audience as behavioral interventions should not be based on a flawed theory.

5 Conclusion

This thesis has provided evidence that a lack of cognitive effort is not a general cause of errors in reasoning. Reasoning errors are associated with both more cognitive effort and less cognitive effort, dependent on the task at hand. In general, cognitive effort has been associated with increased task performance. However, performance on rational reasoning tasks from the heuristics and bias literature is negatively related to subjective cognitive effort experienced on the task. Suggesting that less cognitive effort, not more, is associated with higher performance on these tasks. Thus, providing evidence against the default-interventionist account of bias in reasoning. Furthermore, this indicates that these tasks should not be applied as measures of cognitive effort. The results from this thesis further indicate that NFC, N-TLX, COG-ED, and pupillometry can be applied as measures of cognitive effort, although they have separate strengths and weaknesses. However, the DST in the current version should not be applied as an individual difference measure of cognitive effort. There is a need to develop and validate new cognitive effort measures and further evaluate and cross-validate alternative versions of existing task paradigms.

The results from this thesis largely oppose dual-process models. From the three-stage model of analytic engagement, the concept of conflict detection failure as a source of bias is supported. Further, the dissociation between conflict detection and cognitive decoupling as dissociable sources of bias in reasoning is partially supported. However, the results are mixed regarding competing intuitions as a source of analytic engagement. Importantly, it is not clear that two separate processes are necessary or add explanatory power above single process models of decision-making The thesis highlights the role of the LC-NE system in errors in reasoning. Further, the thesis presents results in line with sequential sampling models and the extensive integration account of bias in reasoning. However, it should be noted that these models were not tested to the same degree as dual-process models. The results supporting sequential sampling models and the extensive integration account should therefore be considered preliminary. However, these models and the methodology applied in this thesis provides a path forward for research on bias and errors in reasoning.

References

Agster, K. L., Mejias-Aponte, C. A., Clark, B. D., & Waterhouse, B. D. (2013). Evidence for a regional specificity in the density and distribution of noradrenergic varicosities in rat cortex. *The Journal of Comparative Neurology*, *521*(10), 2195–2207. https://doi.org/10.1002/cne.23270

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science*, *205*(4412), 1289–1292. https://doi.org/10.1126/science.472746

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Apps, M. A. J., Grima, L. L., Manohar, S., & Husain, M. (2015). The role of cognitive effort in subjective reward devaluation and risky decision-making. *Scientific Reports*, *5*, 16880. https://doi.org/10.1038/srep16880

Arnold, N. R., Bröder, A., & Bayen, U. J. (2015). Empirical validation of the diffusion model for recognition memory and a comparison of parameter-estimation methods. *Psychological Research*, *79*(5), 882–898. https://doi.org/10.1007/s00426-014-0608-y

Arnsten, A. F. (2000). Through the looking glass: Differential noradenergic modulation of prefrontal cortical function. *Neural Plasticity*, *7*(1–2), 133–146. https://doi.org/10.1155/NP.2000.133

Aston-Jones, G., Chen, S., Zhu, Y., & Oshinsky, M. L. (2001). A neural circuit for circadian regulation of arousal. *Nature Neuroscience*, *4*(7), 732–738. https://doi.org/10.1038/89522

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Aston-Jones, G., Foote, S. L., & Bloom, F. E. (1984). Anatomy and physiology of locus coeruleus neurons: Functional implications. *Anatomy and Physiology of Locus Coeruleus Neurons: Functional Implications*, *2*, 92–116.

Aston-Jones, G., Rajkowski, J., & Cohen, J. (1999). Role of locus coeruleus in attention and behavioral flexibility. *Biological Psychiatry*, *46*(9), 1309–1320. https://doi.org/10.1016/S0006-3223(99)00140-7

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3), 241–254. https://doi.org/10.1017/S0140525X07001653

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beatty, J., & Lucero-Wagoner, B. (2000). *The pupillary system*. Cambridge University Press.

Berridge, C. W., & Spencer, R. C. (2016). Differential cognitive actions of norepinephrine a2 and a1 receptor signaling in the prefrontal cortex. *Brain Research*, *1641*(Pt B), 189–196. https://doi.org/10.1016/j.brainres.2015.11.024

Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus–noradrenergic system: Modulation of behavioral state and state-dependent cognitive processes. *Brain Research Reviews*, *42*(1), 33–84. https://doi.org/10.1016/S0165-0173(03)00143-7

Białek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. *Judgment and Decision Making*, *12*(2), 148–167. https://doi.org/10.1017/S1930297500005696

Blaise, M., Marksteiner, T., Krispenz, A., & Bertrams, A. (2021). Measuring Motivation for Cognitive Effort as State. *Frontiers in Psychology*, *12*, 785094. https://doi.org/10.3389/fpsyg.2021.785094

Boag, R. J., Strickland, L., Heathcote, A., Neal, A., Palada, H., & Loft, S. (2023). Evidence accumulation modelling in the wild: Understanding safety-critical decisions. *Trends in Cognitive Sciences*, *27*(2), 175–188. https://doi.org/10.1016/j.tics.2022.11.009

Boersma, F., Wilton, K., Barham, R., & Muir, W. (1970). Effects of arithmetic problem difficulty on pupillary dilation in normals and educable retardates. *Journal of Experimental Child Psychology*, *9*(2), 142–155. https://doi.org/10.1016/0022-0965(70)90079-2

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Bonner, S. E., & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society*, *27*(4), 303–345. https://doi.org/10.1016/S0361-3682(01)00052-6

Bornemann, B., Foth, M., Horn, J., Ries, J., Warmuth, E., Wartenburger, I., & van der Meer, E. (2010). Mathematical cognition: Individual differences in resource allocation. *ZDM Mathematics Education*, *42*(6), 555–567. https://doi.org/10.1007/s11858-010-0253-x

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology*, *66*, 83–113. https://doi.org/10.1146/annurev-psych-010814-015044

Botvinick, M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*(6), 1249–1285. https://doi.org/10.1111/cogs.12126

Botvinick, M., Huffstetler, S., & McGuire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 16–27. https://doi.org/10.3758/CABN.9.1.16

Bouret, S., & Sara, S. J. (2004). Reward expectation, orientation of attention and locus coeruleus-medial frontal cortex interplay during learning. *The European Journal of Neuroscience*, *20*(3), 791–802. https://doi.org/10.1111/j.1460-9568.2004.03526.x

Bouret, S., & Sara, S. J. (2005). Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, *28*(11), 574–582. https://doi.org/10.1016/j.tins.2005.09.002

Braarud, P. Ø. (2021). Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human–machine work. *International Journal of Industrial Ergonomics*, *86*, 103233. https://doi.org/10.1016/j.ergon.2021.103233

Bradshaw, J. L. (1968). Pupil Size and Problem Solving. *Quarterly Journal of Experimental Psychology*, *20*(2), 116–122. https://doi.org/10.1080/14640746808400139

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology*, *40*, 109–131. https://doi.org/10.1146/annurev.ps.40.020189.000545

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, *48*, 94–103. https://doi.org/10.1016/j.chb.2015.01.038

Busemeyer, J. R., Jessup, R. K., Johnson, J. G., & Townsend, J. T. (2006). Building bridges between neural models and complex decision making behaviour. *Neural Networks: The Official Journal of the International Neural Network Society*, *19*(8), 1047–1058. https://doi.org/10.1016/j.neunet.2006.05.043

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459. https://doi.org/10.1037/0033-295X.100.3.432

Bustamante, L. A., Oshinowo, T., Lee, J. R., Tong, E., Burton, A. R., Shenhav, A., Cohen, J. D., & Daw, N. D. (2023). Effort Foraging Task reveals positive correlation between individual differences in the cost of cognitive and physical effort in humans. *Proceedings of the National Academy of Sciences*, *120*(50), e2221510120. https://doi.org/10.1073/pnas.2221510120

Buttenheim, A., Moffitt, R., & Beatty, A. (Eds.). (2023). *Behavioral Economics: Policy Impact and Future Directions*. National Academies Press. https://doi.org/10.17226/26874

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., Feinstein, J., & Jarvis, B. (1996). Dispositional Differences in Cognitive Motivation: The Life and Times of Individuals Varying in Need for Cognition. *Psychological Bulletin*, *119*, 197–253. https://doi.org/10.1037/0033-2909.119.2.197

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Cai, W., Chen, T., Ryali, S., Kochalka, J., Li, C.-S. R., & Menon, V. (2016). Causal Interactions Within a Frontal-Cingulate-Parietal Network During Cognitive Control: Convergent Evidence from a Multisite-Multitask Investigation. *Cerebral Cortex*, *26*(5), 2140–2153. https://doi.org/10.1093/cercor/bhv046

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS Computational Biology*, *17*(3), e1008863. https://doi.org/10.1371/journal.pcbi.1008863

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*(11), 1462–1467. https://doi.org/10.1038/nn.2925

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye Tracking and Pupillometry are Indicators of Dissociable Latent Decision Processes. *Journal of Experimental Psychology: General*, *143*(4), 1476–1488. https://doi.org/10.1037/a0035813

Chandler, D. J., Waterhouse, B. D., & Gao, W.-J. (2014). New perspectives on catecholaminergic regulation of executive circuits: Evidence for independent modulation of prefrontal functions by midbrain dopaminergic and noradrenergic neurons. *Frontiers in Neural Circuits*, *8*, 53. https://doi.org/10.3389/fncir.2014.00053

Chang, W. C., Westbrook, A., Strauss, G. P., Chu, A. O. K., Chong, C. S. Y., Siu, C. M. W., Chan, S. K. W., Lee, E. H. M., Hui, C. L. M., Suen, Y. M., Lo, T. L., & Chen, E. Y. H. (2020). Abnormal cognitive effort allocation and its association with amotivation in first-episode psychosis. *Psychological Medicine*, *50*(15), 2599–2609. https://doi.org/10.1017/S0033291719002769

Chatham, C. H., Frank, M. J., & Munakata, Y. (2009). Pupillometric and behavioral markers of a developmental shift in the temporal dynamics of cognitive control. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 5529–5533. https://doi.org/10.1073/pnas.0810002106

Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature Communications*, *9*(1), 3557. https://doi.org/10.1038/s41467-018-05994-9

Chiew, K. S., & Braver, T. S. (2013). Temporal Dynamics of Motivation-Cognitive Control Interactions Revealed by High-Resolution Pupillometry. *Frontiers in Psychology*, *4*, 15. https://doi.org/10.3389/fpsyg.2013.00015

Chong, T. T.-J., Apps, M., Giehl, K., Sillence, A., Grima, L. L., & Husain, M. (2017). Neurocomputational mechanisms underlying subjective valuation of effort costs. *PLOS Biology*, *15*(2), e1002598. https://doi.org/10.1371/journal.pbio.1002598

Clay, G., Mlynski, C., Korb, F. M., Goschke, T., & Job, V. (2022). Rewarding cognitive effort increases the intrinsic value of mental labor. *Proceedings of the National Academy of Sciences*, *119*(5), e2111785119. https://doi.org/10.1073/pnas.2111785119

Colling, J., Wollschläger, R., Keller, U., Preckel, F., & Fischbach, A. (2022). Need for Cognition and its relation to academic achievement in different learning environments. *Learning and Individual Differences*, *93*, 102110. https://doi.org/10.1016/j.lindif.2021.102110

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. *Biological Psychiatry*, *82*(6), 431–439. https://doi.org/10.1016/j.biopsych.2017.05.017

Cowan, N., Rouder, J. N., Blume, C. L., & Saults, J. S. (2012). Models of verbal working memory capacity: What does it take to make them work? *Psychological Review*, *119*(3), 480–499. https://doi.org/10.1037/a0027791

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent Variable and Latent Structure Models* (pp. 195–224). Lawrence Erlbaum.

Culbreth, A., Westbrook, A., & Barch, D. (2016). Negative symptoms are associated with an increased subjective cost of cognitive effort. *Journal of Abnormal Psychology*, *125*(4), 528–536. https://doi.org/10.1037/abn0000153

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292. https://doi.org/10.1177/1088868313495594

da Silva Castanheira, K., LoParco, S., & Otto, A. R. (2021). Task-evoked pupillary responses track effort exertion: Evidence from task-switching. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(3), 592–606. https://doi.org/10.3758/s13415-020-00843-z

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, *17*(4), 335–350. https://doi.org/10.1080/09548980601004024

de Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, *7*, 10996. https://doi.org/10.1038/ncomms10996

de Chantal, P.-L., Newman, I. R., Thompson, V., & Markovits, H. (2020). Who resists belief-biased inferences? The role of individual differences in reasoning strategies, working memory, and attentional focus. *Memory & Cognition*, *48*(4), 655–671. https://doi.org/10.3758/s13421-019-00998-2

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(5), E618-625. https://doi.org/10.1073/pnas.1317557111 de Gee, J. W., Tsetsos, K., Schwabe, L., Urai, A. E., McCormick, D., McGinley, M. J., & Donner, T. H. (2020). Pupil-linked phasic arousal predicts a reduction of choice bias across species and decision domains. *eLife*, *9*, e54014. https://doi.org/10.7554/eLife.54014

De Neys, W. (2006). Automatic-heuristic and executive-analytic processing during reasoning: Chronometric and dual-task considerations. *Quarterly Journal of Experimental Psychology*, *59*(6), 1070–1100. https://doi.org/10.1080/02724980543000123

De Neys, W. (2018). *Dual process theory 2.0.* (pp. viii, 159). Routledge/Taylor & Francis Group.

De Neys, W. (2020). Rational rationalization and System 2. *The Behavioral and Brain Sciences*, *43*, e34. https://doi.org/10.1017/S0140525X19002048

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Devos, H., Gustafson, K., Ahmadnezhad, P., Liao, K., Mahnken, J. D., Brooks, W. M., & Burns, J. M. (2020). Psychometric Properties of NASA-TLX and Index of Cognitive Activity as Measures of Cognitive Workload in Older Adults. *Brain Sciences*, *10*(12), 994. https://doi.org/10.3390/brainsci10120994

Diederich, A. (2003). MDFT account of decision making under time pressure. *Psychonomic Bulletin & Review*, *10*(1), 157–166. https://doi.org/10.3758/BF03196480

Ding, L., & Gold, J. I. (2012). Neural correlates of perceptual decision making before, during, and after decision commitment in monkey frontal eye field. *Cerebral Cortex*, *22*(5), 1052–1067. https://doi.org/10.1093/cercor/bhr178

DiYanni, C., & Kelemen, D. (2005). Time to get a new mountain? The role of function in children's conceptions of natural kinds. *Cognition*, *97*(3), 327–335. https://doi.org/10.1016/j.cognition.2004.10.002

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., … Donkin, C. (2019). The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. *Psychonomic Bulletin & Review*, *26*(4), 1051–1069. https://doi.org/10.3758/s13423-017-1417-2

Eisenberger, R. (1992). Learned industriousness. *Psychological Review*, *99*(2), 248–267. https://doi.org/10.1037/0033-295X.99.2.248

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. https://doi.org/10.1038/nn.3428

Eldar, E., Cohen, J. D., & Niv, Y. (2016). Amplified selectivity in cognitive processing implements the neural gain model of norepinephrine function. *Behavioral and Brain Sciences*, *39*, e206. https://doi.org/10.1017/S0140525X15001776

Eldar, E., Felso, V., Cohen, J. D., & Niv, Y. (2021). A pupillary index of susceptibility to decision biases. *Nature Human Behaviour*, *5*(5), 653–662. https://doi.org/10.1038/s41562-020-01006-3

Eldar, E., Niv, Y., & Cohen, J. D. (2016). Do You See the Forest or the Tree? Neural Gain and Breadth Versus Focus in Perceptual Processing. *Psychological Science*, *27*(12), 1632–1643. https://doi.org/10.1177/0956797616665578

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *The American Psychologist*, *49*(8), 709–724. https://doi.org/10.1037//0003-066x.49.8.709

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395. https://doi.org/10.3758/BF03193858

Evans, J. St. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. St. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382–389. https://doi.org/10.1080/13546780542000005

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Evans, N. J., & Wagenmakers, E.-J. (2020). Evidence Accumulation Models: Current Limitations and Future Directions. *The Quantitative Methods for Psychology*, *16*(2), 73–90. https://doi.org/10.20982/tqmp.16.2.p073

Fallon, S. J., van der Schaaf, M. E., ter Huurne, N., & Cools, R. (2017). The Neurocognitive Cost of Enhancing Cognition with Methylphenidate: Improved Distractor Resistance but Impaired Updating. *Journal of Cognitive Neuroscience*, *29*(4), 652–663. https://doi.org/10.1162/jocn_a_01065

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective & Behavioral Neuroscience*, *14*(1), 129–146. https://doi.org/10.3758/s13415-013-0236-9

Ferreira, M. B., Soro, J. C., Reis, J., Mata, A., & Thompson, V. A. (2022). When Type 2 Processing Misfires: The Indiscriminate Use of Statistical Thinking about Reasoning Problems. *Journal of Intelligence*, *10*(4), 109. https://doi.org/10.3390/jintelligence10040109

Fisher, G. (2021). A multiattribute attentional drift diffusion model. *Organizational Behavior and Human Decision Processes*, *165*, 167–182. https://doi.org/10.1016/j.obhdp.2021.04.004

Fleischhauer, M., Strobel, A., Enge, S., & Strobel, A. (2013). Assessing implicit cognitive motivation: Developing and testing an Implicit Association Test to measure need for cognition. *European Journal of Personality*, *27*(1), 15–29. https://doi.org/10.1002/per.1841

Foote, S. L., Bloom, F. E., & Aston-Jones, G. (1983). Nucleus locus ceruleus: New evidence of anatomical and physiological specificity. *Physiological Reviews*, *63*(3), 844–914. https://doi.org/10.1152/physrev.1983.63.3.844

Foote, S. L., Freedman, R., & Oliver, A. P. (1975). Effects of putative neurotransmitters on neuronal activity in monkey auditory cortex. *Brain Research*, *86*(2), 229–242. https://doi.org/10.1016/0006-8993(75)90699-X

Forstmann, B., Brown, S., Dutilh, G., Neumann, J., & Wagenmakers, E.-J. (2010). The neural substrate of prior information in perceptual decision making: A model-based analysis. *Frontiers in Human Neuroscience*, *4*. https://www.frontiersin.org/articles/10.3389/fnhum.2010.00040

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Friedman, N. P., & Miyake, A. (2017). Unity and Diversity of Executive Functions: Individual Differences as a Window on Cognitive Structure. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Froböse, M. I., Swart, J. C., Cook, J. L., Geurts, D. E. M., den Ouden, H. E. M., & Cools, R. (2018). Catecholaminergic modulation of the avoidance of cognitive control. *Journal of Experimental Psychology: General*, *147*(12), 1763–1781. https://doi.org/10.1037/xge0000523

Gailliot, M. T., & Baumeister, R. F. (2007). The physiology of willpower: Linking blood glucose to self-control. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *11*(4), 303–327. https://doi.org/10.1177/1088868307303030

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., Brewer, L. E., & Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology*, *92*(2), 325–336. https://doi.org/10.1037/0022-3514.92.2.325

Garner, K. G., & Dux, P. E. (2015). Training conquers multitasking costs by dividing task representations in the frontoparietal-subcortical system. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(46), 14372–14377. https://doi.org/10.1073/pnas.1511423112

Gärtner, A., Grass, J., Wolff, M., Goschke, T., Strobel, A., & Strobel, A. (2021). No relation of Need for Cognition to basic executive functions. *Journal of Personality*, *89*(6), 1113–1125. https://doi.org/10.1111/jopy.12639

Germar, M., Albrecht, T., Voss, A., & Mojzisch, A. (2016). Social conformity is due to biased stimulus processing: Electrophysiological and diffusion analyses. *Social Cognitive and Affective Neuroscience*, *11*(9), 1449–1459. https://doi.org/10.1093/scan/nsw050

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. https://doi.org/10.1126/science.aac6076

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

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: A Journal of the American Psychological Society*, *8*(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 513–525. https://doi.org/10.1037/0096-1523.14.3.513

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

Gold, J. I., Law, C.-T., Connolly, P., & Bennur, S. (2008). The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. *Journal of Neurophysiology*, *100*(5), 2653–2668. https://doi.org/10.1152/jn.90629.2008

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Gold, J. M., Kool, W., Botvinick, M. M., Hubzin, L., August, S., & Waltz, J. A. (2015). Cognitive effort avoidance and detection in people with schizophrenia. *Cognitive, Affective & Behavioral Neuroscience*, *15*(1), 145–154. https://doi.org/10.3758/s13415-014-0308-5

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*(4), 457–461. https://doi.org/10.1111/j.1469-8986.1996.tb01071.x

Greene, J. D. (2014). Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics*, *124*(4), 695–726. https://doi.org/10.1086/675875

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9). https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hartig, F. (2022). *DHARMa—Residual Diagnostics for HierARchical Models* (Version 0.4.6) [R package]. http://florianhartig.github.io/DHARMa/

Hayes, T., & Usami, S. (2020). Factor Score Regression in the Presence of Correlated Unique Factors. *Educational and Psychological Measurement*, *80*(1), 5–40. https://doi.org/10.1177/0013164419854492

Hess, E. H., & Polt, J. M. (1964). Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Hess, E. H., Seltzer, A. L., & Shlien, J. M. (1965). Pupil response of hetero- and homosexual males to pictures of men and women: A pilot study. *Journal of Abnormal Psychology*, *70*(3), 165–168. https://doi.org/10.1037/h0021978

Hill, B. D., Foster, J. D., Elliott, E. M., Shelton, J. T., McCain, J., & Gouvier, Wm. D. (2013). Need for cognition is related to higher general intelligence, fluid intelligence, and crystallized intelligence, but not working memory. *Journal of Research in Personality*, *47*(1), 22–25. https://doi.org/10.1016/j.jrp.2012.11.001

Hockey, G. R. (1997). Compensatory control in the regulation of human performance under stress and high workload; a cognitive-energetical framework. *Biological Psychology*, *45*(1–3), 73–93. https://doi.org/10.1016/s0301-0511(96)05223-4

Hockey, G. R. J. (2011). A motivational control theory of cognitive fatigue. In P. L. Ackerman (Ed.), *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications* (pp. 167–187). American Psychological Association. https://doi.org/10.1037/12343-008

Holroyd, C. B. (2016). The waste disposal problem of effortful control. In T. S. Braver (Ed.), *Motivation and cognitive control* (pp. 235–260). Routledge/Taylor & Francis Group.

Horan, W. P., Reddy, L. F., Barch, D. M., Buchanan, R. W., Dunayevich, E., Gold, J. M., Marder, S. R., Wynn, J. K., Young, J. W., & Green, M. F. (2015). Effort-Based Decision-Making Paradigms for Clinical Trials in Schizophrenia: Part 2—External Validity and Correlates. *Schizophrenia Bulletin*, *41*(5), 1055–1065. https://doi.org/10.1093/schbul/sbv090

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory* (pp. x, 422). Appleton-Century.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Inzlicht, M., & Schmeichel, B. J. (2012). What Is Ego Depletion? Toward a Mechanistic Revision of the Resource Model of Self-Control. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(5), 450–463. https://doi.org/10.1177/1745691612454134

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The Effort Paradox: Effort Is Both Costly and Valued. *Trends in Cognitive Sciences*, *22*(4), 337–349. https://doi.org/10.1016/j.tics.2018.01.007

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Johnson, D. A. (1971). Pupillary responses during a short-term memory task: Cognitive processing, arousal, or both? *Journal of Experimental Psychology*, *90*(2), 311–318. https://doi.org/10.1037/h0031562

Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review*, *112*(4), 841–861. https://doi.org/10.1037/0033-295X.112.4.841

Jones, B. E., Halaris, A. E., McIlhany, M., & Moore, R. Y. (1977). Ascending projections of the locus coeruleus in the rat. I. Axonal transport in central noradrenaline neurons. *Brain Research*, *127*(1), 1–21. https://doi.org/10.1016/0006-8993(77)90377-8

Jones, B. E., & Yang, T. Z. (1985). The efferent projections from the reticular formation and the locus coeruleus studied by anterograde and retrograde axonal transport in the rat. *The Journal of Comparative Neurology*, *242*(1), 56–92. https://doi.org/10.1002/cne.902420105

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*, *89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Just, M. A., Carpenter, P. A., & Miyake, A. (2003). Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science*, *4*(1–2), 56–88. https://doi.org/10.1080/14639220210159735

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*(4), 407–424. https://doi.org/10.1017/S1930297500005271

Kahan, D. M. (2015). The Politically Motivated Reasoning Paradigm. *Emerging Trends in Social & Behavioral Sciences, Forthcoming*. https://papers.ssrn.com/abstract=2703011

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Kahneman, D., & Wright, P. (1971). Changes of pupil size and rehearsal strategies in a short-term memory task. *The Quarterly Journal of Experimental Psychology*, *23*(2), 187–196. https://doi.org/10.1080/14640747108400239

Kelemen, D. (1999). Function, goals and intention: Children’s teleological reasoning about objects. *Trends in Cognitive Sciences*, *3*(12), 461–468. https://doi.org/10.1016/S1364-6613(99)01402-3

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*(4), 1074–1083. https://doi.org/10.1037/a0030399

Keuken, M. C., Van Maanen, L., Bogacz, R., Schäfer, A., Neumann, J., Turner, R., & Forstmann, B. U. (2015). The subthalamic nucleus during decision-making with multiple alternatives. *Human Brain Mapping*, *36*(10), 4041–4052. https://doi.org/10.1002/hbm.22896

Kim, E. H., Morse, A., & Zingales, L. (2006). What Has Mattered to Economics Since 1970. *Journal of Economic Perspectives*, *20*(4), 189–202. https://doi.org/10.1257/jep.20.4.189

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–17. https://doi.org/10.1017/S0140525X00041157

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology. General*, *143*(1), 131–141. https://doi.org/10.1037/a0031048

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, *2*(12), Article 12. https://doi.org/10.1038/s41562-018-0401-9

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*(4), 665–682. https://doi.org/10.1037/a0020198

Krajbich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, *29*, 6–11. https://doi.org/10.1016/j.copsyc.2018.10.008

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), Article 10. https://doi.org/10.1038/nn.2635

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*(1), Article 1. https://doi.org/10.1038/ncomms8455

Krajbich, I., Lu, D., Camerer, C., & Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in Psychology*, *3*, 193. https://doi.org/10.3389/fpsyg.2012.00193

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857. https://doi.org/10.1073/pnas.1101328108

Kramer, A.-W., Van Duijvenvoorde, A. C. K., Krabbendam, L., & Huizenga, H. M. (2021). Individual differences in adolescents' willingness to invest cognitive effort: Relation to need for cognition, motivation and cognitive capacity. *Cognitive Development*, *57*. https://doi.org/10.1016/j.cogdev.2020.100978

Kramer, S. E., Lorens, A., Coninx, F., Zekveld, A. A., Piotrowska, A., & Skarzynski, H. (2013). Processing load during listening: The influence of task characteristics on the pupil response. *Language and Cognitive Processes*, *28*(4), 426–442. https://doi.org/10.1080/01690965.2011.642267

Kreis, I., Moritz, S., & Pfuhl, G. (2020). Objective Versus Subjective Effort in Schizophrenia. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01469

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109. https://doi.org/10.1037/a0020762

Kuhn, M. (2015). caret: Classification and Regression Training. *Astrophysics Source Code Library*, ascl:1505.003.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *The Behavioral and Brain Sciences*, *36*(6), 10.1017/S0140525X12003196. https://doi.org/10.1017/S0140525X12003196

Laeng, B., & Alnaes, D. (2019). Pupillometry. In C. Klein & U. Ettinger (Eds.), *Eye Movement Research: An Introduction to its Scientific Foundations and Applications* (pp. 449–502). Springer International Publishing. https://doi.org/10.1007/978-3-030-20085-5_11

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A Window to the Preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

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. https://doi.org/10.1037/0003-066X.43.6.431

Leite, F. P., & Ratcliff, R. (2010). Modeling Reaction Time and Accuracy of Multiple-Alternative Decisions. *Attention, Perception & Psychophysics*, *72*(1), 246–273. https://doi.org/10.3758/APP.72.1.246

Lerche, V., von Krause, M., Voss, A., Frischkorn, G. T., Schubert, A.-L., & Hagemann, D. (2020). Diffusion modeling and intelligence: Drift rates show both domain-general and domain-specific relations with intelligence. *Journal of Experimental Psychology. General*, *149*(12), 2207–2249. https://doi.org/10.1037/xge0000774

Lerche, V., & Voss, A. (2017). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, *81*(3), 629–652. https://doi.org/10.1007/s00426-016-0770-5

Lerche, V., & Voss, A. (2019). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychological Research*, *83*(6), 1194–1209. https://doi.org/10.1007/s00426-017-0945-8

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, *49*(2), 513–537. https://doi.org/10.3758/s13428-016-0740-2

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence*, *30*(1), 81–108. https://doi.org/10.1016/0004-3702(86)90068-8

Lin, H., Pennycook, G., & Rand, D. G. (2023). Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, *230*, 105312. https://doi.org/10.1016/j.cognition.2022.105312

Lopez-Gamundi, P., & Wardle, M. C. (2018). The cognitive effort expenditure for rewards task (C-EEfRT): A novel measure of willingness to expend cognitive effort. *Psychological Assessment*, *30*(9), 1237–1248. https://doi.org/10.1037/pas0000563

MacDonald, E., Kobilka, B. K., & Scheinin, M. (1997). Gene targeting—Homing in on alpha 2-adrenoceptor-subtype function. *Trends in Pharmacological Sciences*, *18*(6), 211–219. https://doi.org/10.1016/s0165-6147(97)01063-8

Mækelæ, M. J., Klevjer, K., Westbrook, A., Eby, N. S., Eriksen, R., & Pfuhl, G. (2023). Is it cognitive effort you measure? Comparing three task paradigms to the Need for Cognition scale. *PLOS ONE*, *18*(8), e0290177. https://doi.org/10.1371/journal.pone.0290177

Mækelæ, M. J., & Pfuhl, G. (2019). Deliberate reasoning is not affected by language. *PLOS ONE*, *14*(1), e0211428. https://doi.org/10.1371/journal.pone.0211428

Markovits, H., de Chantal, P.-L., Brisson, J., Dubé, É., Thompson, V., & Newman, I. (2021). Reasoning strategies predict use of very fast logical reasoning. *Memory & Cognition*, *49*(3), 532–543. https://doi.org/10.3758/s13421-020-01108-3

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, *17*(1), 11–17. https://doi.org/10.3758/bf03199552

Martiny-Huenger, T., Bieleke, M., Doerflinger, J., Stephensen, M. B., & Gollwitzer, P. M. (2021). Deliberation decreases the likelihood of expressing dominant responses. *Psychonomic Bulletin & Review*, *28*(1), 139–157. https://doi.org/10.3758/s13423-020-01795-8

Massar, S. A. A., Libedinsky, C., Weiyan, C., Huettel, S. A., & Chee, M. W. L. (2015). Separate and overlapping brain areas encode subjective value during delay and effort discounting. *NeuroImage*, *120*, 104–113. https://doi.org/10.1016/j.neuroimage.2015.06.080

Mathot, S. (2018). Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition*, *1*(1), 16. https://doi.org/10.5334/joc.18

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

Mathôt, S., & Vilotijević, A. (2023). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behavior Research Methods*, *55*(6), 3055–3077. https://doi.org/10.3758/s13428-022-01957-7

McBurney-Lin, J., Lu, J., Zuo, Y., & Yang, H. (2019). Locus coeruleus-norepinephrine modulation of sensory processing and perception: A focused review. *Neuroscience and Biobehavioral Reviews*, *105*, 190–199. https://doi.org/10.1016/j.neubiorev.2019.06.009

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159–188. https://doi.org/10.1037/0096-3445.114.2.159

McGaughy, J., Ross, R. S., & Eichenbaum, H. (2008). Noradrenergic, but not cholinergic, deafferentation of prefrontal cortex impairs attentional set-shifting. *Neuroscience*, *153*(1), 63–71. https://doi.org/10.1016/j.neuroscience.2008.01.064

McGuire, J. T., & Botvinick, M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(17), 7922–7926. https://doi.org/10.1073/pnas.0910662107

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Mittner, M., Hawkins, G. E., Boekel, W., & Forstmann, B. U. (2016). A Neural Model of Mind Wandering. *Trends in Cognitive Sciences*, *20*(8), 570–578. https://doi.org/10.1016/j.tics.2016.06.004

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Molden, D. C., Hui, C. M., Scholer, A. A., Meier, B. P., Noreen, E. E., D’Agostino, P. R., & Martin, V. (2012). Motivational versus metabolic effects of carbohydrates on self-control. *Psychological Science*, *23*(10), 1137–1144. https://doi.org/10.1177/0956797612439069

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/s1364-6613(03)00028-7

Moore, R. Y., & Bloom, F. E. (1979). Central catecholamine neuron systems: Anatomy and physiology of the norepinephrine and epinephrine systems. *Annual Review of Neuroscience*, *2*, 113–168. https://doi.org/10.1146/annurev.ne.02.030179.000553

Mridha, Z., de Gee, J. W., Shi, Y., Alkashgari, R., Williams, J., Suminski, A., Ward, M. P., Zhang, W., & McGinley, M. J. (2021). Graded recruitment of pupil-linked neuromodulation by parametric stimulation of the vagus nerve. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21730-2

Murphy, P. R., Moort, M. L. van, & Nieuwenhuis, S. (2016). The Pupillary Orienting Response Predicts Adaptive Behavioral Adjustment after Errors. *PLOS ONE*, *11*(3), e0151763. https://doi.org/10.1371/journal.pone.0151763

Murphy, P. R., Vandekerckhove, J., & Nieuwenhuis, S. (2014). Pupil-Linked Arousal Determines Variability in Perceptual Decision Making. *PLOS Computational Biology*, *10*(9), e1003854. https://doi.org/10.1371/journal.pcbi.1003854

Musslick, S., Cohen, J. D., & Shenhav, A. (2018). Estimating the costs of cognitive control from task performance: Theoretical validation and potential pitfalls. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018*, 798–803. https://collaborate.princeton.edu/en/publications/estimating-the-costs-of-cognitive-control-from-task-performance-t

Musslick, S., Dey, B., Ozcimder, K., Patwary, M. A., Willke, T. L., & Cohen, J. D. (2016, August 8). Parallel processing capability versus efficiency of representation in neural networks. *15th*

Neural Comput. Psychol. Workshop. https://doi.org/A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior.10.13140/RG.2.2.30076.54407

Nagai, T., Satoh, K., Imamoto, K., & Maeda, T. (1981). Divergent projections of catecholamine neurons of the locus coeruleus as revealed by fluorescent retrograde double labeling technique. *Neuroscience Letters*, *23*(2), 117–123. https://doi.org/10.1016/0304-3940(81)90027-6

Nagase, A. M., Onoda, K., Foo, J. C., Haji, T., Akaishi, R., Yamaguchi, S., Sakai, K., & Morita, K. (2018). Neural Mechanisms for Adaptive Learned Avoidance of Mental Effort. *Journal of Neuroscience*, *38*(10), 2631–2651. https://doi.org/10.1523/JNEUROSCI.1995-17.2018

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–520. https://doi.org/10.1007/s00213-006-0502-4

Noguchi, T., & Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological Review*, *125*(4), 512–544. https://doi.org/10.1037/rev0000102

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11*(6), 988–1010. https://doi.org/10.3758/BF03196730

Otero, I., Salgado, J. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence*, *90*, 101614. https://doi.org/10.1016/j.intell.2021.101614

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*(1), 46–59. https://doi.org/10.1002/hbm.20131

Padmala, S., & Pessoa, L. (2011). Reward Reduces Conflict by Enhancing Attentional Control and Biasing Visual Cortical Processing. *Journal of Cognitive Neuroscience*, *23*(11), 3419–3432. https://doi.org/10.1162/jocn_a_00011

Pan, W., Geng, H., Zhang, L., Fengler, A., Frank, M., ZHANG, R.-Y., & Chuan-Peng, H. (2022). *A Hitchhiker's Guide to Bayesian Hierarchical Drift-Diffusion Modeling with dockerHDD*. https://doi.org/10.31234/osf.io/6uzga

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian Stochastic Modelling in Python. *Journal of Statistical Software*, *35*(4), 1–81.

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019). The transdiagnostic structure of mental effort avoidance. *Scientific Reports*, *9*(1), 1689. https://doi.org/10.1038/s41598-018-37802-1

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In *Handbook of research methods in personality psychology* (pp. 224–239). The Guilford Press.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335–346. https://doi.org/10.1016/j.cognition.2012.03.003

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*(1), 101–106. https://doi.org/10.1016/j.cognition.2012.04.004

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015a). Everyday Consequences of Analytic Thinking. *Current Directions in Psychological Science*, *24*(6), 425–432. https://doi.org/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. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., Fugelsang, J. A., Koehler, D. J., & Thompson, V. A. (2016). Commentary: Rethinking fast and slow based on a critique of reaction-time reverse inference. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01174

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, *19*(3), 528–534. https://doi.org/10.3758/s13423-012-0249-3

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 544–554. https://doi.org/10.1037/a0034887

Persson, E., Andersson, D., Koppel, L., Västfjäll, D., & Tinghög, G. (2021). A preregistered replication of motivated numeracy. *Cognition*, *214*, 104768. https://doi.org/10.1016/j.cognition.2021.104768

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Poe, G. R., Foote, S., Eschenko, O., Johansen, J. P., Bouret, S., Aston-Jone, G., Harley, C. W., Manahan-Vaughan, D., Weinshenker, D., Valentino, R., Berridge, C., Chandler, D. J., Waterhouse, B., & Sara, S. J. (2020). Locus coeruleus: A new look at the blue spot. *Nature Reviews. Neuroscience*, *21*(11), 644–659. https://doi.org/10.1038/s41583-020-0360-9

Poock, G. K. (1973). Information processing vs pupil diameter. *Perceptual and Motor Skills*, *37*(3), 1000–1002. https://doi.org/10.1177/003151257303700363

Preuschoff, K., 't Hart, B., & Einhauser, W. (2011). Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. *Frontiers in Neuroscience*, *5*, 115.

Puveendrakumaran, P., Fervaha, G., Caravaggio, F., & Remington, G. (2020). Assessing analytic and intuitive reasoning using the cognitive reflection test in young patients with schizophrenia. *Psychiatry Research*, *284*, 112683. https://doi.org/10.1016/j.psychres.2019.112683

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Research Bulletin*, *35*(5–6), 607–616. https://doi.org/10.1016/0361-9230(94)90175-9

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, *14*, 170–178.

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85,* 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.

Reddy, L. F., Horan, W. P., Barch, D. M., Buchanan, R. W., Dunayevich, E., Gold, J. M., Lyons, N., Marder, S. R., Treadway, M. T., Wynn, J. K., Young, J. W., & Green, M. F. (2015). Effort-Based Decision-Making Paradigms for Clinical Trials in Schizophrenia: Part 1—Psychometric Characteristics of 5 Paradigms. *Schizophrenia Bulletin*, *41*(5), 1045–1054. https://doi.org/10.1093/schbul/sbv089

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*(1), 13289. https://doi.org/10.1038/ncomms13289

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370. https://doi.org/10.1037/0033-295X.108.2.370

Roxin, A., & Ledberg, A. (2008). Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. *PLoS Computational Biology*, *4*(3), e1000046. https://doi.org/10.1371/journal.pcbi.1000046

Rubinstein, A. (2007). Instinctive and Cognitive Reasoning: A Study of Response Times*. *The Economic Journal*, *117*(523), 1243–1259. https://doi.org/10.1111/j.1468-0297.2007.02081.x

Rudolph, J., Greiff, S., Strobel, A., & Preckel, F. (2018). Understanding the link between need for cognition and complex problem solving. *Contemporary Educational Psychology*, *55*, 53–62. https://doi.org/10.1016/j.cedpsych.2018.08.001

Samuels, E. R., & Szabadi, E. (2008). Functional neuroanatomy of the noradrenergic locus coeruleus: Its roles in the regulation of arousal and autonomic function part I: principles of functional organisation. *Current Neuropharmacology*, *6*(3), 235–253. https://doi.org/10.2174/157015908785777229

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. https://doi.org/10.1016/j.cognition.2017.12.004

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: The locus coeruleus mediates cognition through arousal. *Neuron*, *76*(1), 130–141. https://doi.org/10.1016/j.neuron.2012.09.011

Sara, S. J., & Segal, M. (1991). Plasticity of sensory responses of locus coeruleus neurons in the behaving rat: Implications for cognition. *Progress in Brain Research*, *88*, 571–585. https://doi.org/10.1016/s0079-6123(08)63835-2

Sayalı, C., & Badre, D. (2019). Neural systems of cognitive demand avoidance. *Neuropsychologia*, *123*, 41–54. https://doi.org/10.1016/j.neuropsychologia.2018.06.016

Sayalı, C., & Badre, D. (2021). Neural systems underlying the learning of cognitive effort costs. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(4), 698–716. https://doi.org/10.3758/s13415-021-00893-x

Schaefer, T., Ferguson, J. B., Klein, J. A., & Rawson, E. B. (1968). Pupillary responses during mental activities. *Psychonomic Science*, *12*(4), 137–138. https://doi.org/10.3758/BF03331236

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Schwarz, L. A., Miyamichi, K., Gao, X. J., Beier, K. T., Weissbourd, B., DeLoach, K. E., Ren, J., Ibanes, S., Malenka, R. C., Kremer, E. J., & Luo, L. (2015). Viral-genetic tracing of the input-output organization of a central noradrenaline circuit. *Nature*, *524*(7563), 88–92. https://doi.org/10.1038/nature14600

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science*, *249*(4971), 892–895. https://doi.org/10.1126/science.2392679

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791–806. https://doi.org/10.1016/j.neuron.2013.10.047

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222. https://doi.org/10.1037/0033-2909.134.2.207

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, *40*, 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Shenhav, A., Rand, D. G., & Greene, J. D. (2017). The relationship between intertemporal choice and following the path of least resistance across choices, preferences, and beliefs. *Judgment and Decision Making*, *12*(1), 1–18. https://doi.org/10.1017/S1930297500005209

Sidarus, N., Palminteri, S., & Chambon, V. (2019). Cost-benefit trade-offs in decision-making and learning. *PLOS Computational Biology*, *15*(9), e1007326. https://doi.org/10.1371/journal.pcbi.1007326

Simon, H. A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and Probability* (pp. 15–18). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-20568-4_5

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews. Cognitive Science*, *5*(6), 679–692. https://doi.org/10.1002/wcs.1323

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, *32*(2), 135–168. https://doi.org/10.1016/0022-2496(88)90043-0

Smullyan, R. M. (1978). *What is the name of this book?: The riddle of Dracula and other logical puzzles*. Prentice-Hall.

Solomon, R. L. (1948). The influence of work on behavior. *Psychological Bulletin*, *45*(1), 1–40. https://doi.org/10.1037/h0055527

Soubelet, A., & Salthouse, T. A. (2017). Does need for cognition have the same meaning at different ages? *Assessment*, *24*(8), 987–998. https://doi.org/10.1177/1073191116636449

Spaniol, J., Voss, A., & Grady, C. L. (2008). Aging and emotional memory: Cognitive mechanisms underlying the positivity effect. *Psychology and Aging*, *23*(4), 859–872. https://doi.org/10.1037/a0014218

Spencer, R. C., & Berridge, C. W. (2019). Receptor and circuit mechanisms underlying differential procognitive actions of psychostimulants. *Neuropsychopharmacology*, *44*(10), 1820–1827. https://doi.org/10.1038/s41386-019-0314-y

Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (1st ed., pp. 55–88). Oxford University PressOxford. https://doi.org/10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K. E. (2009b). *What Intelligence Tests Miss: The Psychology of Rational Thought*. Yale University Press.

Stanovich, K. E. (2016). The comprehensive assessment of rational thinking. *Educational Psychologist*, *51*, 23–34. https://doi.org/10.1080/00461520.2015.1125787

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & Toplak, M. E. (2012). Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, *11*(1), 3–13. https://doi.org/10.1007/s11299-011-0093-6

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342–357. https://doi.org/10.1037/0022-0663.89.2.342

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Strobel, A., Fleischhauer, M., Enge, S., & Strobel, A. (2015). Explicit and implicit need for cognition and bottom-up/top-down attention allocation. *Journal of Research in Personality, 55,* 10–13. https://doi.org/10.1016/j.jrp.2014.11.002

Strobel, A., Wieder, G., Paulus, P. C., Ott, F., Pannasch, S., Kiebel, S. J., & Kührt, C. (2020). Dispositional cognitive effort investment and behavioral demand avoidance: Are they related? *PloS One*, *15*(10), e0239817. https://doi.org/10.1371/journal.pone.0239817

Takeuchi, T., Duszkiewicz, A. J., Sonneborn, A., Spooner, P. A., Yamasaki, M., Watanabe, M., Smith, C. C., Fernández, G., Deisseroth, K., Greene, R. W., & Morris, R. G. M. (2016). Locus coeruleus and dopaminergic consolidation of everyday memory. *Nature*, *537*(7620), Article 7620. https://doi.org/10.1038/nature19325

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review*, *120*(1), 1–38. https://doi.org/10.1037/a0030776

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness* (pp. x, 293). Yale University Press.

Thompson, V. A., & Markovits, H. (2021). Reasoning strategy vs cognitive capacity as predictors of individual differences in reasoning performance. *Cognition, 217,* 104866. https://doi.org/10.1016/j.cognition.2021.104866

Thompson, V. A., & Newman, I. R. (2020). Working memory, autonomy, and dual process theories: A roadmap. In S. Elqayam, I. Douven, J. S. B. T. Evans, & N. Cruz (Eds.), *Logic and Uncertainty in the Human Mind*. Routledge.

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. St. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General*, *147*, 945–961. https://doi.org/10.1037/xge0000457

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making, 11*(1), 99–113. https://doi.org/10.1017/S1930297500007622

Thomson, K. S., & Oppenheimer, D. M. (2022). The “Effort Elephant” in the Room: What Is Effort, Anyway? *Perspectives on Psychological Science*, *17*(6), 1633–1652. https://doi.org/10.1177/17456916211064896

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*(7), 1275–1289. https://doi.org/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*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Tran, T., Spilka, M. J., Ruiz, I., & Strauss, G. P. (2022). Implicit cognitive effort monitoring impairments are associated with expressive negative symptoms in schizophrenia. *Schizophrenia Research*, *248*, 14–20. https://doi.org/10.1016/j.schres.2022.07.006

Trippas, D., & Handley, S. J. (2018). The parallel processing model of belief bias: Review and extensions. In W. De Neys (Ed.), *Dual process theory 2.0.* (pp. 28–46). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315204550-3

Trippas, D., Pennycook, G., Verde, M. F., & Handley, S. J. (2015). Better but still biased: Analytic cognitive style and belief bias. *Thinking & Reasoning*, *21*(4), 431–445. https://doi.org/10.1080/13546783.2015.1016450

Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition*, *45*(4), 539–552. https://doi.org/10.3758/s13421-016-0680-1

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*(2), 179–205. https://doi.org/10.1037/a0036137

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences*, *109*(24), 9659–9664. https://doi.org/10.1073/pnas.1119569109

Tubbs-Cooley, H. L., Mara, C. A., Carle, A. C., & Gurses, A. P. (2018). The NASA Task Load Index as a measure of overall workload among neonatal, paediatric and adult intensive care nurses. *Intensive & Critical Care Nursing*, *46*, 64–69. https://doi.org/10.1016/j.iccn.2018.01.004

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206. https://doi.org/10.1016/j.neuroimage.2013.01.048

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage*, *128*, 96–115. https://doi.org/10.1016/j.neuroimage.2015.12.030

Turner, B. M., Van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review*, *122*(2), 312–336. https://doi.org/10.1037/a0038894

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Uematsu, A., Tan, B. Z., Ycu, E. A., Cuevas, J. S., Koivumaa, J., Junyent, F., Kremer, E. J., Witten, I. B., Deisseroth, K., & Johansen, J. P. (2017). Modular organization of the brainstem noradrenaline system coordinates opposing learning states. *Nature Neuroscience*, *20*(11), 1602–1611. https://doi.org/10.1038/nn.4642

United Nations. (2021). *Verified | #Pledgetopause*. https://shareverified.com/pledge-to-pause/

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective & Behavioral Neuroscience*, *16*(4), 601–615. https://doi.org/10.3758/s13415-016-0417-4

Unsworth, N., & Robison, M. K. (2017). A locus coeruleus-norepinephrine account of individual differences in working memory capacity and attention control. *Psychonomic Bulletin & Review*, *24*(4), 1282–1311. https://doi.org/10.3758/s13423-016-1220-5

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*(1), 14637. https://doi.org/10.1038/ncomms14637

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The Role of Locus Coeruleus in the Regulation of Cognitive Performance. *Science*, *283*(5401), 549–554. https://doi.org/10.1126/science.283.5401.549

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295x.108.3.550

Usher, M., & McClelland, J. L. (2004). Loss Aversion and Inhibition in Dynamical Models of Multialternative Choice. *Psychological Review*, *111*, 757–769. https://doi.org/10.1037/0033-295X.111.3.757

Usher, M., Tsetsos, K., Lagnado, D., & Yu, E. (2013). Dynamics of decision-making: From evidence accumulation to preference and belief. *Frontiers in Psychology, 4*, 785.

van de Ven, N., Gilovich, T., & Zeelenberg, M. (2010). Delay, Doubt, and Decision: How Delaying a Choice Reduces the Appeal of (Descriptively) Normative Options. *Psychological Science*, *21*(4), 568–573. https://doi.org/10.1177/0956797610363546

van der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., Kramer, J., Warmuth, E., Heekeren, H. R., & Wartenburger, I. (2010). Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology*, *47*(1), 158–169. https://doi.org/10.1111/j.1469-8986.2009.00884.x

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Van Gestel, L. C., Adriaanse, M. A., & De Ridder, D. T. D. (2021). Do nudges make use of automatic processing? Unraveling the effects of a default nudge under type 1 and type 2 processing. *Comprehensive Results in Social Psychology*, *5*(1–3), 4–24. https://doi.org/10.1080/23743603.2020.1808456

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*(2), 208–256. https://doi.org/10.3758/bf03212980

Vassena, E., Krebs, R. M., Silvetti, M., Fias, W., & Verguts, T. (2014). Dissociating contributions of ACC and vmPFC in reward prediction, outcome, and choice. *Neuropsychologia*, *59*, 112–123. https://doi.org/10.1016/j.neuropsychologia.2014.04.019

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences*, *2*(3), 169–194. https://doi.org/10.1023/A:1022371901259

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Viglione, A., Mazziotti, R., & Pizzorusso, T. (2023). From pupil to the brain: New insights for studying cortical plasticity through pupillometry. *Frontiers in Neural Circuits*, *17*. https://www.frontiersin.org/articles/10.3389/fncir.2023.1151847

Vogel, T. A., Savelson, Z. M., Otto, A. R., & Roy, M. (2020). Forced choices reveal a trade-off between cognitive effort and physical pain. *eLife*, *9*, e59410. https://doi.org/10.7554/eLife.59410

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220. https://doi.org/10.3758/BF03196893

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A Diffusion Model Account of Criterion Shifts in the Lexical Decision Task. *Journal of Memory and Language*, *58*(1), 140–159. https://doi.org/10.1016/j.jml.2007.04.006

Wang, C., Trongnetrpunya, A., Samuel, I. B. H., Ding, M., & Kluger, B. M. (2016). Compensatory Neural Activity in Response to Cognitive Fatigue. *Journal of Neuroscience*, *36*(14), 3919–3924. https://doi.org/10.1523/JNEUROSCI.3652-15.2016

Wang, M., Ramos, B. P., Paspalas, C. D., Shu, Y., Simen, A., Duque, A., Vijayraghavan, S., Brennan, A., Dudley, A., Nou, E., Mazer, J. A., McCormick, D. A., & Arnsten, A. F. T. (2007). Alpha2A-adrenoceptors strengthen working memory networks by inhibiting cAMP-HCN channel signaling in prefrontal cortex. *Cell*, *129*(2), 397–410. https://doi.org/10.1016/j.cell.2007.03.015

Waterhouse, B. D., & Navarra, R. L. (2019). The locus coeruleus-norepinephrine system and sensory signal processing: A historical review and current perspectives. *Brain Research*, *1709*, 1–15. https://doi.org/10.1016/j.brainres.2018.08.032

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. https://doi.org/10.1037/a0012842

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective & Behavioral Neuroscience*, *15*(2), 395–415. https://doi.org/10.3758/s13415-015-0334-y

Westbrook, A., & Braver, T. S. (2016). Dopamine Does Double Duty in Motivating Cognitive Effort. *Neuron*, *89*(4), 695–710. https://doi.org/10.1016/j.neuron.2015.12.029

Westbrook, A., Frank, M. J., & Cools, R. (2021). A mosaic of cost-benefit control over cortico-striatal circuitry. *Trends in Cognitive Sciences*, *25*(8), 710–721. https://doi.org/10.1016/j.tics.2021.04.007

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*(7), e68210. https://doi.org/10.1371/journal.pone.0068210

Westbrook, A., Lamichhane, B., & Braver, T. (2019). The Subjective Value of Cognitive Effort is Encoded by a Domain-General Valuation Network. *Journal of Neuroscience*, *39*(20), 3934–3947. https://doi.org/10.1523/JNEUROSCI.3071-18.2019

Westbrook, A., van den Bosch, R., Määttä, J. I., Hofmans, L., Papadopetraki, D., Cools, R., & Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, *367*(6484), 1362–1366. https://doi.org/10.1126/science.aaz5891

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 14. https://doi.org/10.3389/fninf.2013.00014

Wiehler, A., Branzoli, F., Adanyeguh, I., Mochel, F., & Pessiglione, M. (2022). A neuro-metabolic account of why daylong cognitive work alters the control of economic decisions. *Current Biology: CB*, *32*(16), 3564-3575.e5. https://doi.org/10.1016/j.cub.2022.07.010

Yang, X., & Krajbich, I. (2023). A dynamic computational model of gaze and choice in multi-attribute decisions. *Psychological Review*, *130*(1), 52–70. https://doi.org/10.1037/rev0000350

Yu, A. J., & Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge. *Trends in Hearing*, *22*. https://doi.org/10.1177/2331216518777174

Zerna, J., Scheffel, C., Kührt, C., & Strobel, A. (2023). Need for Cognition is associated with a preference for higher task load in effort discounting. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-44349-3

Zipf, G. K. (1949). *Human behavior and the principle of least effort* (pp. xi, 573). Addison-Wesley Press.

Paper 1

RESEARCH ARTICLE

Is it cognitive effort you measure? Comparing three task paradigms to the Need for Cognition scale

Martin Jensen Mækelæ[1], **Kristoffer Klevjer**[1], **Andrew Westbrook**[2], **Noah S. Eby**[3], **Rikke Eriksen**[1], **Gerit Pfuhl**[1,4]*

1 Department of Psychology, UiT–The Arctic University of Norway, Tromsø, Norway, **2** Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, United States of America, **3** Department of Neurology, University of Washington, Seattle, WA, United States of America, **4** Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

* gerit.pfuhl@ntnu.no

Citation: Mækelæ MJ, Klevjer K, Westbrook A, Eby NS, Eriksen R, Pfuhl G (2023) Is it cognitive effort you measure? Comparing three task paradigms to the Need for Cognition scale. PLoS ONE 18(8): e0290177. https://doi.org/10.1371/journal.pone.0290177

Editor: Mark Fenton-O'Creevy, The Open University, UNITED KINGDOM

Received: April 12, 2023

Accepted: August 2, 2023

Published: August 17, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0290177

Data Availability Statement: Data can be found at the Open Science Framework, https://osf.io/dywt4/.

Funding: The author(s) received no specific funding for this work.

Abstract

Measuring individual differences in cognitive effort can be elusive as effort is a function of motivation and ability. We report six studies (N = 663) investigating the relationship of Need for Cognition and working memory capacity with three cognitive effort measures: demand avoidance in the Demand Selection Task, effort discounting measured as the indifference point in the Cognitive Effort Discounting paradigm, and rational reasoning score with items from the heuristic and bias literature. We measured perceived mental effort with the NASA task load index. The three tasks were not correlated with each other (all r's < .1, all *p*'s > .1). Need for Cognition was positively associated with effort discounting (r = .168, p < .001) and rational reasoning (r = .176, p < .001), but not demand avoidance (r = .085, p = .186). Working memory capacity was related to effort discounting (r = .185, p = .004). Higher perceived effort was related to poorer rational reasoning. Our data indicate that two of the tasks are related to Need for Cognition but are also influenced by a participant's working memory capacity. We discuss whether any of the tasks measure cognitive effort.

Introduction

Laziness is built deep into our nature (Kahneman, 2011, p. 39)

People tend to choose the least demanding line of action, famously formulated as the "Law of least work" [1]. Although originally applied to physical effort, it also applies to effort in the cognitive domain [2]. The underlying assumption is that there is a cost associated with cognitive effort [3, 4]. The nature of this cost is uncertain [5] but brain imaging studies have shown that increased cognitive effort reduces activity in the reward network [6–8]. It has been proposed that cognitive effort depends on a cost-benefit analysis to find an optimal balance of expenditure [3, 9–12]. However, cognitive effort is everything but well operationalized [13]. Effort has been described as the use of executive functions, use of attention, workload or computational constraints [13].

Competing interests: The authors have declared that no competing interests exist.

Proposed explanations for cognitive effort costs include resource limits and computational costs [4, 14–18], metabolic costs or accumulation of by-products [19], and opportunity costs [4, 12]. Assertions of cognitive effort costs and minimization have been proposed to be implicated in a range of fields e.g., behavioral economics [20, 21], executive functions [22], linguistics [23], and judgment and decision-making [24]. Effort is often inferred from the outcome, i.e., answering intuitively is effortless whereas analytically is effortful. As such, effort is often assumed but rarely validated. Not least because of a missing operationalization and its tight relationship with motivation and cognitive ability. This paper is beyond solving the effort problem [13]. Instead, we present six experiments where we compare three measures of cognitive effort against the benchmark Need for Cognition scale (NCS) and report the subjective task demands of each task with the NASA task load index (N-TLX).

There are well-established individual differences in the willingness to engage in cognitively effortful tasks. Those individual differences can be reliably measured with the Need for Cognition Scale [25, 26]. Still, behavioral paradigms measuring cognitive effort are useful for investigating actualized cognitive effort expenditure, decision-making, developmental trajectories and neural underpinnings. Additionally, concerns about the reliability and validity of self-report motivate the use of behavioral paradigms to complement self-report instruments [27]. Behavioral tasks can be combined with physiological measures and used across the lifespan. Accordingly, a range of tasks have been developed to measure cognitive effort spent in a task. We here focus on cognitive effort, though physical and perceptual effort tasks have been developed too [for a review see e.g., 28, 29].

One strand of research uses computerized tasks for measuring choices between cognitively more or less demanding options. Here, choice patterns are seen as an indication of cognitive effort costs or preferences to avoid cognitive effort [30–34]. Another strand of research gauges typical cognitive effort expenditure by using tasks that require cognitively demanding deliberate processing to answer correctly [35–38]. These approaches differ in numerous ways and show partly opposing results, also when used in clinical samples [32, 39–43]. It is therefore of importance to assess to what degree the paradigms measure the same “cognitive effort” construct.

Task paradigms for measuring cognitive effort

Rationality battery. Task performance on rational reasoning tasks (RQ) is an alternative way of measuring thinking disposition or “cognitive miserliness” [35, 44–46]. Thinking disposition is proposed to be on a spectrum with one end being the preference for using computationally more demanding mechanisms for solving tasks, known as an analytic thinking disposition. On the other end of the spectrum is a preference for cognitive shortcuts, namely an intuitive thinking disposition. An intuitive thinking disposition is prone to rely more on heuristics, which can serve to reduce cognitive effort [24]. Task performance on rational reasoning tasks is proposed to depend on using more cognitively demanding mechanisms and avoiding overreliance on heuristic responses (avoiding “miserly information processing”) [47]. Suppression of intuitive but wrong answers requires cognitive control [38]. Individual differences have previously been noticed in tasks measuring deliberate reasoning [48]. Toplak et al. [37] showed that the cognitive reflection task, assesses both the ability and willingness to perform cognitive work. However, recent work has questioned whether normative responding is effortful [49–53]. Performance may depend on cognitive ability, not effort [35–37, 54]. Firstly, normative responding can be as fast as heuristic responding [50, 53]. Secondly, the CRT has been shown to correlate highly with numerical tasks [55] and deliberation and rational thinking are highly correlated with cognitive ability [51]. Note, the rationality battery used here is

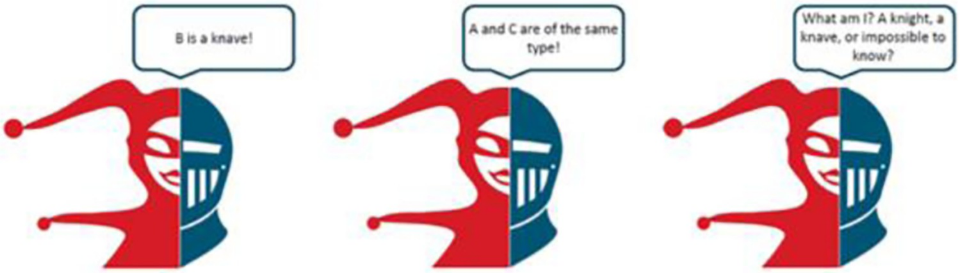

B is a knave!
A and C are of the same type!
What am I? A knight, a knave, or impossible to know?

Fig 1. Example of a task from the rationality battery. 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 says: "B is a knave." B says: "A and C are of the same type." What is C?.

https://doi.org/10.1371/journal.pone.0290177.g001

more than the cognitive reflection test (Fig 1). Such items have been shown to correlate positively with the CRT [37, 56] and the Need for Cognition scale [56]. The CRT has been shown to be positively associated with the Need for Cognition scale too [38], but see [55].

Demand selection task. Evidence to support cognitive effort minimization or demand avoidance was shown with the Demand Selection Task paradigm (DST, Fig 2) by Kool, McGuire et al. [30]. In this task, participants make either parity or magnitude judgements for numerical digits. Effort demands are manipulated by the frequency of task shifts: one line of action (high demand) has more frequent task shifts, thus increasing effort demand [57]. DST can be considered an implicit measure of cognitive effort or demand avoidance as participants are not informed of the demands of the tasks or given any incentive to choose high or low demand lines of action. However, several participants detect the demand manipulation, and some evidence suggests this leads to increased effort avoidance [41].

Cognitive effort discounting paradigm. Westbrook et al. [32] were able to quantify the individual differences in effort costs with the Cognitive Effort Discounting Paradigm (COGED, Fig 3). In this paradigm, participants make repeated choices between performing a low demand working memory task (1-back) for a small reward or performing a high demand working memory task for a larger reward (n-back, n being 2, 3, 4, 5, or 6). The reward for the low demand task is titrated in response to participants' choices with the aim to find a subjective indifference point between the low demand and high demand option. The COGED thereby quantifies the subjective monetary discounting due to cognitive effort costs across multiple demand levels. Given that task load levels and offer amounts are all explicit, COGED is an explicit cognitive effort measure. Participants experience the effort demand for each load level prior to making choices between explicit monetary offers.

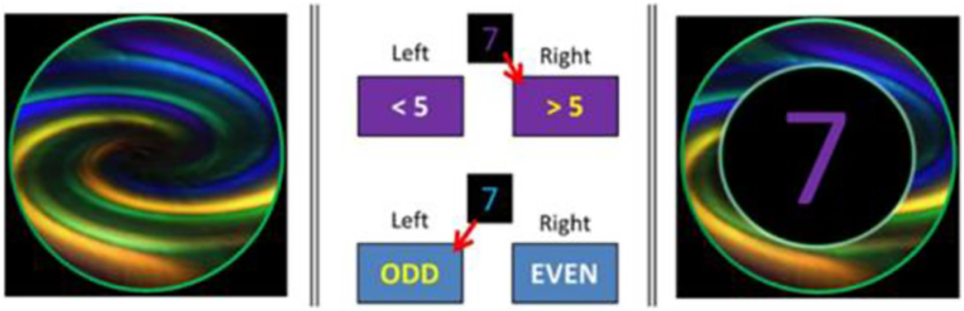


Fig 2. Schematic illustration of the demand selection task. In this trial correct responding is by pressing the right mouse button. Note: Participants saw the rules at the beginning and had to remember them during the test blocks.

https://doi.org/10.1371/journal.pone.0290177.g002

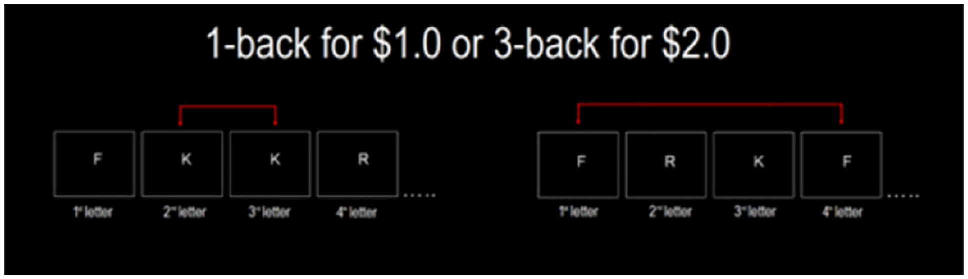


Fig 3. Schematic illustration of the choice phase of the cognitive effort discounting task. Note: Participants saw the n-back instructions during training, in the choice phase they were asked to play 1-back vs n-back and the value for 1-back was titrated either up (if n-back chosen) or down (if 1-back chosen).

https://doi.org/10.1371/journal.pone.0290177.g003

The current studies

An outstanding question is whether cognitive effort can reliably be measured. If the three tasks share a common latent construct, that of willingness or propensity to exert cognitive effort, we would expect that all three tasks are related to a measure of enjoying and engaging in cognitively demanding tasks. These individual trait differences in thinking disposition can reliably be measured with the Need for Cognition Scale [25]. The scale has good internal consistency, test-retest reliability, and measurement invariance [26]. People who score high on Need for Cognition (cognizers) seek, evaluate and integrate multiple relevant sources before arriving at an opinion. People who score low on Need for Cognition (cogmisers) tend to use less demanding cognitive processes [25, 58]. Cognizers may value effort whereas cogmisers may avoid effort [59].

If the three behavioral tasks index task-invariant cognitive effort, we expect them to be positively correlated with each other and with the NCS. If the behavioral tasks are not related to each other, the propensity to exert cognitive effort is task-specific. Each task might still be related to the NCS.

Previous work has already shown a positive relationship between effort discounting in the COGED and NCS, and between rational reasoning and NCS [32, 35, 36, 60], but no relationship between demand avoidance in the DST and NCS [42]. It is also not known whether effort discounting is related to rational reasoning and whether cognitive demand avoidance is related to effort discounting and or rational reasoning. Notably, both COGED and DST have been used as measures of cognitive effort in clinical and developmental research [32, 41, 61–64].

We measured test-retest reliability for the DST, COGED and rationality battery. Previously, Stagnaro, Pennycook et al. [65] have shown good test-retest reliability (r = .806) for the cognitive reflection test. Strobel et al. [2020] found questionable test-retest reliability for the DST (ρ = .61). To the best of our knowledge test-retest reliability of the COGED has not been shown.

Finally, subjective effort, which may deviate from objective effort [e.g., 66], can be assessed with the NASA task load index (N-TLX) [67]. Westbrook et al. [32] found increasing subjective ratings of mental and physical effort, temporal demand, failure rate, effort demand and frustration for increasing working memory load levels in the COGED but did not assess it for the choice phase. Here, we report six studies from two independent labs investigating the relationship between demand avoidance in the DST, effort discounting in the COGED, rational reasoning score, NCS and N-TLX. We controlled for working memory capacity by using the n-back performance assessed in the practice phase of the COGED. We also assessed test-retest reliability for the rationality battery, DST and COGED. We report two-sided and non-corrected p-values per study and the mean effect size based on meta-analysis approach. The

S1 File contains details to methods and results of the studies as well as an alternative analysis and plots.

Study 1–6

Participants: Study 1 and 2

All study procedures including informed consent were approved by the Institutional Review Board at Washington University in Saint Louis. The studies were conducted in 2013. Participants provided written informed consent.

In study 1 all participants were undergraduate students at Washington University in St. Louis, USA (N = 76, 49 female). The mean age was 21.43 (range 18 to 32 years). In study 2 participants were 91 undergraduate students (47 female, 35 male) at Washington University in St. Louis. The mean age was 23.62 (range 18 to 40 years). Two participants each in study 1 and 2 were excluded due to very bad performance in the n-back task (negative d'). Final sample size for study 1 is N = 74, and for study 2 is N = 80. There was no missing data.

Participants: Study 3–6

The studies were approved by the institutional review board at the Department of Psychology, UiT–The Arctic University of Norway. The studies were conducted in 2018–2022. All participants provided written informed consent on paper and informed consent (online study 6), respectively.

In study 3 participants were 102 (62 female, 25 male, 15 unknown) undergraduate psychology students at UiT–The Arctic University of Norway and testing was over two sessions. 65 completed both sessions, 82 completed the NCS and the rational reasoning battery, 78 completed the COGED and rational reasoning battery, 63 completed the NCS and COGED. The mean age was 22.6 (range 20 to 38 years).

In study 4 we recruited 40 participants (27 female, range 18 to 37 years). 34 were students at UiT–The Arctic University of Norway, three were full-time workers, and three were high school students. All participants completed both testing sessions. One participant performed randomly in the Demand Selection task and was excluded, i.e., N = 39. Another participant had missing data for the rational reasoning battery.

In study 5 all participants were students (non-psychology) at UiT–The Arctic University of Norway (N = 45, 27 female), mean age was 23.35 (range 18 to 37 years). There was no missing data.

In study 6 participants (*M* = 26.64 years) were recruited from two psychology courses at UiT–The Arctic University of Norway (N = 91, 67 female, 22 male, 2 non-binary; range 19 to 38 years) and from Prolific (prolific.co) (N = 227, 113 female, 110 male, 4 non-binary, range 18 to 62 years). Three participants aborted the choice phase in the COGED and were excluded for parts of the data analysis.

Materials

Effort discounting (alterations across studies in brackets). The COGED task was administered through E-prime 2.0 (Psychology Software Tools, Inc., Sharpsburg, PA) in study 1+2 and through Inquisit (Millisecond.com) in study 3–6. The task started with a practice phase of the n-back task [68]. Participants played all load levels for three runs (six levels in study 1 and 2, four levels in study 3–6).

Study 1 and 2

All runs consisted of 64 items (consonants, presented in Courier New font, font size 24). Items were presented on screen for 1.5 seconds, during which participants could respond. After 1.5 seconds the items were replaced by a fixation cross. The inter-trial interval was 3.5 seconds. Participants were given feedback about % of targets and % of non-targets correct. Feedback of "Good job!" was given if both scores were above 50% or "Please try harder!" if not. From this phase d' was calculated as an index of working memory capacity (see below).

In the discounting procedure participants were offered to play n = 1 for a small reward or n > 1 for a larger reward. Participants were offered six choices for each load level. The amount for the higher offer (n > 1) was always $2. The reward amount for the lower offer (n = 1) started at $1 and was adjusted up if participants chose the high offer and was adjusted down if participants chose the low offer. Each time a choice was made, the reward amount was adjusted to half as much as in the previous choice. After the last choice (six choices in total), the amount was adjusted to $0.015. The final amount was taken as the participants' subjective indifference point. Participants played five load levels and made six choices for each level, yielding 30 choices in total. To ensure choices reflected participants' preferences, they were told that one of the choices would be selected for them to repeat 10 more times and they would be paid for each repetition. Further, they were told that payment was contingent on maintaining effort, but not on performance. Effort would be monitored by "behavioral clues". All participants completed their randomly selected offer four times and were paid the associated amount.

Study 3–6

The first phase consisted of five runs per n-back level (2, 3, & 4), each run with five target trials (responding would be a hit), and 10+N non-target trials (responding would be a false alarm) in a pseudo-random sequence. Each trial lasted 2.5 s, and in each trial, participants were presented with a stimulus (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 the 2 seconds had to either respond (press 'A' on the keyboard) or not to 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 the discounting procedure for 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. The discounting amounts were identical to study 1 and 2. In study 3 and 5 participants were informed that they would not receive extra money, thereby eliminating external reward as motivation. In study 4 they could earn a bonus on top of the show-up fee. In study 6 performance in the discounting procedure phase had to be at least 80% (previously 80% for 1-back but at least 100% of that from the practice phase for 2-back, 3-back and 4-back) to count as success. Participants could earn vouchers (students) or a bonus (Prolific). The bonus was related to the amount earned in the discounting phase of the task.

The Average Indifference Point (AIP) across all load levels is the cognitive effort discounting measure used for the bi-variate correlations and regression analysis.

Cognitive demand avoidance

We used an exact replication of Experiment 3 in Kool et al. (2010). The task was administered on a computer, using MatLab 2018a (The MathWorks, MATLAB, Version 9.4, 2018), with Psychophysics Toolbox 3 extension [69–71]. The task starts with a training phase where participants complete two different tasks. Participants are presented with a number (between 1 and 9, excluding 5). The number can be either blue or yellow. The color of the number signaled the task required on that trial. If the number is blue, participants must decide if the number is higher or lower than 5. If the number is yellow, then participants must decide if the number is odd or even. Participants indicate their choice by clicking on the right or left side of a computer mouse. During the training phase (60 trials), participants received feedback on their performance. None of the participants had to redo the training phase. In the main task, participants see two colorful balls on screen (they appear along an invisible circle at an angular distance of 45 degrees). The location of the balls changes between runs but is stable throughout a run. Participants must sample from each option but are told they can stay with one if they develop a preference. There are eight runs with 75 trials in each run (600 in total). There is one high demand option (ball) where the task switches with a probability of 0.9, and there is a low demand option where the probability of task switch is 0.1. Task instructions were available in paper format in case participants forgot the rules. Demand avoidance is quantified in terms of the proportion selection of the high demand decks (ball)–thus a demand avoidant participant would score between 0 and .5 and preferring the low demand deck, respectively.

In study 6 the abridged version, i.e., four rounds and 300 trials in total, was used [61].

Rational reasoning

In study 2 we used the 18 items scale from Toplak, West [37]. This scale includes the original 3-item Cognitive Reflection Test, measuring individual differences in detecting errors and overriding an initial intuitive response [48]. The remaining 15 items were problems from the heuristics and biases literature: two-sample size problems, two gambler's fallacy problems, regression to the mean, a base rate problem, a covariation detection problem, one Bayesian reasoning problem, one conjunction fallacy problem, a denominator neglect problem, a methodological reasoning problem, a probability matching problem, a sunk cost fallacy problem, one outcome bias problem, and a framing problem. Correct answers were scored as 1, incorrect as 0. Total composite score, the rationality quotient RQ, ranged between 0 and 18.

In study 3, 4 and 5 we used 14 items from the heuristics and biases literature. We used items 2–7 from the Cognitive Reflection Test [35], one fully disjunctive reasoning problem "the marriage problem" [72], one probability matching task [73], one probability estimation task "the bus problem" [74], one making sense of medical results problem [75], one Bayesian reasoning problem [76], one covariation detection problem [77], one knight and knave problem [78], one conditional reasoning problem [79]. Correct answers were scored as 1, incorrect as 0. Total composite score ranged between 0 and 14.

In study 6 we used 12 items. These were items 2–7 from the Cognitive Reflection Test [35], one fully disjunctive reasoning problem, "the marriage problem" [72], one knight and knave problem [78], one conditional reasoning problem [79], one covariation problem [37], one base rate problem [36], one making sense of medical results problem [75]. We calculated the proportion of correct items, i.e., the score ranged from 0 (no item correct) to 1 (all items correct).

Thinking disposition was measured with the 18-item Need for Cognition Scale (NCS) [80]. An example item is *"I prefer complex to simple problems"*. The 18 items are rated on a 5-point Likert scale from 1 = "*Extremely uncharacteristic of me*" to 5 = "*Extremely characteristic of me*". Total score range is from 18 to 90.

Working memory capacity was measured with the d' from the n-back portion of the COGED task. Here, responding to a previously seen stimulus at the n-th position is a hit, not responding is a miss. Responding too early or too late is a false alarm, not responding to an incorrect letter is a correct rejection. We calculated d' from signal detection theory, i.e., d' = z

(H)–z(FA) where z(H) and z(FA) are the z transforms of hit rate and false alarm rate, respectively. The larger d' the better is a participant's working memory capacity. Study 1 and 2: In the O-Span task participants have to remember sequentially presented words and solve simple, interspersed math problems. The length of the sequence reproduced error-free is used as the maximal working memory capacity.

Perceived effort. We asked participants to rate their perceived effort, both mental, physical, temporal, as well as performance, overall effort and frustration using the NASA task load index (N-TLX) [67, 81]. Rating was on a visual analogue scale ranging from 1 = very low to 20 = very high.

Procedure

Study 1 and 2: Participants were paid 10$ per hour for their participation, and they could earn additional money based on their choice in the COGED. Participants received their payment at the end of the testing session. Testing was completed individually at Washington University in St. Louis.

Study 1: Order of the tasks was: DST, COGED, NCS, O-Span (not reported). Usual participation time was approximately two hours.

Study 2: Order of the tasks was: rational reasoning problems, COGED, NCS, O-Span (not reported). Usual participation time was approximately two hours.

Study 3: Testing took place over two separate sessions in small groups in a computer pool at the campus (UiT). Students took part for course credit and received no monetary compensation. Students could choose to partake in only one session. On day 1, 82 students took part, and were tested on COGED and N-TLX$_{\text{COGED}}$. On day 2, approximately 3 weeks later, 84 students took the Rational reasoning items, N-TLX$_{\text{RQ}}$, and NCS. 65 students took part in both test sessions. Participants could withdraw or indicate on the consent form that they do not permit to use their data for research, which was once the case. Each session including debriefing and took approximately 1 hour.

Study 4: All participants were tested individually at UiT–The Arctic University of Norway. All participants completed a second testing session between four and eight weeks after the first testing session. Day 1 task order was; DST, Rational reasoning, Bullshit receptivity scale (not included in analysis), NCS, Effort expenditure for rewards task (EEfRT, [82], not included in analysis) and N-TLX$_{\text{EEfRT}}$ (not included). Day 2 task order was; DST, NCS, Handgrip [83], not included in analysis), COGED, and N-TLX$_{\text{COGED}}$. Participants received a voucher with a fixed amount of 200 NOK (approx. $25) for participation, plus between 50 and 150 NOK depending on task performance in the COGED and EEfRT.

Study 5: Participants were tested individually at UiT–The Arctic University of Norway. All participants received a voucher worth 400 NOK after completing two days of testing (day 2 involved eye tracking, not included here). Each testing session lasted approximately between 1.5 and 2 hours. Relevant is only the first test session. Task order for day 1 was: DST, rational reasoning task, NCS, Teleological reasoning (not reported here).

Study 6: The study was done fully online. Participants read the informed consent and then were randomly assigned to one of the six orders for the three tasks (COGED, DST, rationality reasoning). After each task, they filled out the N-TLX. The NCS was always presented at the end (Prolific sample) and the day after for the UiT students. The tasks were implemented in Inquisit (Millisecond.com). The NCS for UiT participants was implemented in

Qualtrics. Duration was 50–60 min. UiT students received course credit and could earn vouchers and Prolific participants at least £8 for participation plus bonus payment for good performance.

None of the studies was preregistered. All analyses are thus exploratory. With the present sample sizes (minimum is N = 402), we were able to detect correlations of at least r = .177 at an alpha level of .05 and power of .95. Data analysis was done in R [84].

Results across the six studies

Fig 4 presents the descriptive data for rational reasoning (as percentage of maximum score), demand avoidance (choice of high demand option), cognitive effort discounting (average indifference point), NCS, d', and subjective mental effort rating across the six studies (for study 6 N-TLX values have been converted from the 20-point scale to the 100-point scale). Differences between the six studies are reported in S1 File, including a comparison between lab vs online studies.

We performed bi-variate Pearson product-moment correlations per study. Based on the study-wise correlation coefficients we calculated the mean effect size by using the meta for package [85]. The correlation coefficients are shown in Table 1 and the scatterplots per study in S1 File. An alternative analysis using z-scored values and Bayes Factor is provided in the S1 File.

As can be seen from Table 1, we found a significant positive association between NCS and cognitive effort discounting, and between NCS and rational reasoning score. However, there was no association between demand avoidance in the DST and NCS. This is consistent with previous work. Importantly, there were no significant associations between rational reasoning, cognitive effort discounting or demand avoidance. Thus, the results show that none of the three behavioral tasks measuring cognitive effort are related to each other. Rational reasoning and effort discounting were related to working memory capacity (d').

We next performed three linear mixed regressions to assess whether demand avoidance, rational reasoning or cognitive effort discounting (outcome) was predicted by Need for

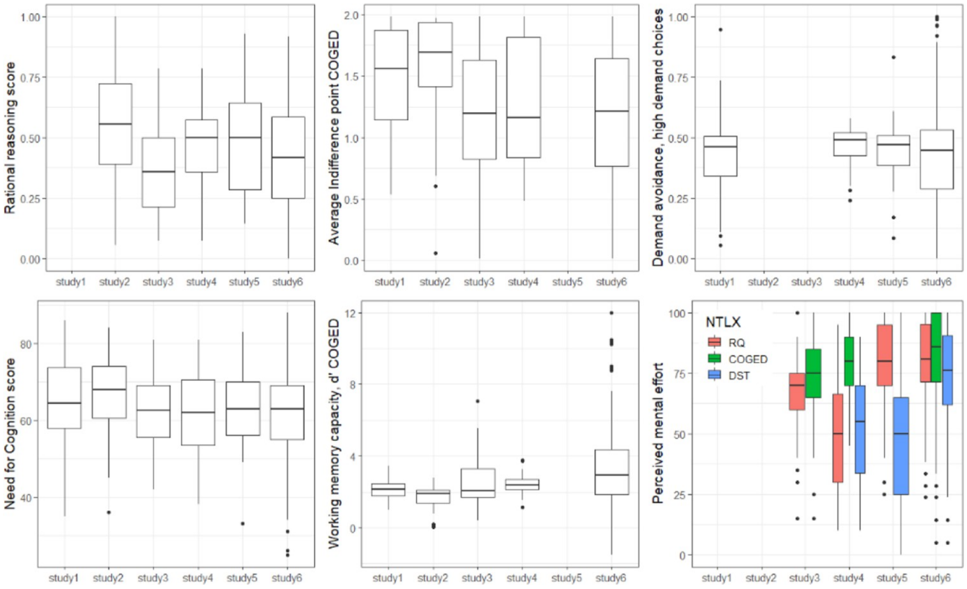


Fig 4. Descriptive data (box plots) for the main outcome variables per study.

https://doi.org/10.1371/journal.pone.0290177.g004

Table 1. Pearson's Correlations per study and overall effect size including confidence intervals and p-value.

Correlations/ study	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	mean r	upper CI	lower CI	p
NCS–d'	0.015	0.091	-0.096	0.180		0.033	0.042	0.117	-0.048	0.208
NCS—IP	0.296	0.115	0.090	0.071		0.177	0.168	0.248	0.086	<0.001
NCS—High Demand	0.060			0.421	-0.190	0.071	0.085	0.265	-0.101	0.186
NCS—RQ		0.171	0.007	0.405	0.225	0.184	0.176	0.270	0.079	<0.001
d'—IP	0.075	0.411	0.094	0.286		0.107	0.185	0.313	0.050	0.004
d'—High demand	0.173			0.106		-0.043	0.040	0.184	-0.107	0.298
d'—RQ		0.501	0.208	0.226		0.257	0.304	0.429	0.166	<0.001
IP—High demand	0.074			0.258		0.038	0.063	0.158	-0.033	0.098
IP—RQ		0.192	-0.002	-0.016		0.067	0.070	0.157	-0.017	0.058
High demand—RQ				0.248	0.076	0.007	0.037	0.15	-0.062	0.233

https://doi.org/10.1371/journal.pone.0290177.t001

Cognition, working memory capacity, any of the other two tasks, or perceived mental effort from the N-TLX, i.e., in each of the three tasks there were five predictors and data was nested within study.

For rational reasoning the fixed effects NCS (t(353) = 3.681, p = .0003), d' (t(353) = 4.797, p < .001) and perceived mental effort (t(353) = -2.518, p = .0124) were significant. Thus, the less effortful the task was perceived, the higher the NCS and the better working memory capacity, the higher was the rational reasoning score.

For demand avoidance none of the fixed effects was significant (all p's > .128). Thus, neither NCS, working memory capacity, perceived mental effort, rational reasoning score or effort discounting were related to the proportion of high demand choices.

For cognitive effort discounting the fixed effect NCS (t(353) = 3.016, p = .0028) was significant. Working memory capacity was not (p = .051), nor any of the other three predictors.

Reliability

In study 4 we assessed test-retest reliability of the DST. Demand avoidance showed acceptable internal consistency on Day 1 (Cronbach's α = .71), but poor internal consistency on Day 2 (Cronbach's α = .52), and poor reliability across the two testing sessions (r = 0.537, *p* < 0.001).

In two adjacent experiments (see S1 File) we assessed test-retest reliability of a rationality battery (similar items to study 3–6) and the choice pattern in the COGED (average indifference points). Test-retest reliability for the rationality items was good, i.e., Pearson's r(83) = 0.70. Test-retest reliability for effort discounting was good, r(25) = 0.804 (average over three test sessions, session 1 with 2: r = 0.789, session 2 with 3: r = 0.819).

General discussion

In six studies we did not find that two common cognitive effort tasks, COGED and DST, as well as items from the problem solving and reasoning literature (rationality battery) were related to each other. However, the COGED and the rational reasoning score were positively correlated with the Need for Cognition score and working memory capacity. This was not the case for the Demand Selection task.

COGED is an explicit incentivized choice task between n-back levels. In contrast, the DST requires detecting which of the decks has fewer rule changes and thereby less effortful. Rational reasoning is not necessarily effortful for participants with high cognitive ability [53]. Thus, there are good reasons which may explain why the three tasks do not relate to each other. Still, of the three tasks the COGED and rational reasoning score had positive correlations with

Need for Cognition, i.e., less effort discounting and better rational reasoning was associated with a thinking disposition for engaging in cognitively effortful tasks, replicating previous studies [32, 35, 36, 60]. Relatedly, better working memory, i.e., how good a participant is in the n-back task, was associated with less effort discounting and also a higher rational reasoning score. The latter is in line with a range of studies finding that cognitive ability predicts performance in heuristic and bias tasks [36, 50]. Neither Need for Cognition nor working memory were related to demand avoidance in the DST. This might be due to the implicit nature of the task or perceptual preferences [30]. Note though, that on average participants did prefer the low demand deck, showing that the task does capture a preference for demand avoidance (Fig 4).

Interestingly, and in line with the smart intuitor account, participants scoring well on the rational reasoning items perceived the task as less effortful. Less effort discounting leads to performing harder n-back trials in the task, i.e., task demand becomes higher, thus participants engaging in least effort discounting perceived the task as more effortful. There was no relation between perceived effort and demand avoidance in the DST.

Regarding test-retest reliability, the demand selection task was not reliable, replicating Strobel, Wieder et al. [43]. The rationality items were reasonable reliable and the COGED had good test-retest reliability.

Rational reasoning items may not measure cognitive effort

Rational reasoning tasks have been used as a convenient, fast and implicit measure of successfully engaging in deliberate reasoning. Indeed, those scoring high on the Need for Cognition scale do perform better on these tasks. Remarkable is also its good test-retest reliability [63]. However, a range of studies question the assumption that performing well on those items is effortful [86–88]. Deliberation can still be effortful, but the items commonly used, also in our studies, may not require deliberation but can be solved by intuition [53]. Researchers should be mindful that performance is dependent on sufficient analytical and reflective abilities yet to be properly defined [89]. Despite task performance being linked to multiple real-world outcomes [90], we caution the use of rational reasoning items to gauge cognitive effort.

Cognitive effort discounting measures cognitive effort

COGED is a behavioral economic approach to assess cognitive effort discounting of monetary rewards. It is a useful tool for explicitly assessing cognitive effort expenditure and cognitive effort costs. The task was subjectively rated as the most mentally demanding task in our studies. COGED is based on the n-back, a well-established working memory paradigm with parametrically varying cognitive load. Thus, a strength of the COGED paradigm is that performance level is adjusted to a participant's ability and performance in the practice phase. In addition, the measure can provide an estimate of working memory, which is convenient as this allows for correction of cognitive ability which is a confounding variable with most cognitive effort measures. By providing feedback through presenting d' after each round, participant may base their choice of n-back level on this feedback. Participants may prefer levels where they performed better. However, high performing individuals might find the 1-back boring, particularly after engaging in higher levels [91]. This could be mitigated by offering e.g., 2 vs 3-back choice options. COGED might be influenced by individual differences in reward sensitivity, as individuals high in Need for Cognition are less sensitive to rewards [92], but see [7]. This underlines the importance of disentangling intrinsic and extrinsic motivation. Notably, real-world academic achievement was not related to effort discounting in adolescents [93] but intrinsic motivation assessed with a questionnaire was. We recommend measuring Need for Cognition or Motivation for Cognitive effort [94] to regress out intrinsic motivation.

Demand avoidance may measure cognitive effort but not reliably

The implicit nature of the DST makes it appealing, however the implicit nature may also limit the tasks' predictive capacity as the task is subject to choices being influenced by factors such as side- and color preferences and also whether or not demand differences were perceived in the first place. In addition, those who detect the demand manipulation show higher demand avoidance compared to those who have not, making the task more into a game to discover the least effortful strategy [30]. The DST showed low test-retest reliability, replicating the finding of [43]. The task was not related to NCS, COGED, or rational reasoning. Given that on average participants did avoid the high demand deck, it is surprising to not find a significant relation with Need for Cognition. For future studies we recommend using a modification of the DST, varying the effort level by changing the frequency of rule changes between rounds [6] and use forced trials to gauge reliable switch costs [30]. Switch costs may index cognitive flexibility and thereby allow to assess relative effort, similar to performance being based on d' in the COGED.

Limitations and strength

Our samples are mostly students, cautioning generalizability beyond young, healthy, well-educated participants. Since the tasks can feel quite repetitive, a subset of participants might have become bored. We did not inquire about participants' level of boredom. The studies used working memory capacity based on the practice phase in the COGED (n-back task) for individual differences in cognitive abilities. The COGED used in study 1 and 2 does differ from the COGED used in study 3 to 6. We did not measure reward sensitivity or liking of challenges [95]. We limited the comparison to these three tasks, not including the Cognitive Effort Expenditure for Rewards task [96] as we were not aware of the task when starting our studies. This paradigm is an incentivized version of the DST.

Our study is the first to compare three common paradigms for measuring cognitive effort. The results replicated across various samples (psychology undergraduates, non-psychology undergraduates, non-students and students recruited at Prolific) and whether instructions were individually, in groups or solely on screen (online testing). Task-specific effects cannot explain why the three tasks do not relate to each other. However, theoretically, demand avoidance in the DST has to be discovered, rational reasoning has been shown to be intuitive for high performers, and effort discounting is reward sensitive. Future studies should carefully manipulate only one of the aspects the three tasks differ on, to identify which component reflects best individual differences in cognitive effort. Cognitive effort may depend on differences in cognitive ability, intrinsic- and extrinsic motivation, reward sensitivity, task automaticity, and effort costs [5]. Using Thomson and Oppenheimer's framework [13], we have not touched on all levels of analysis, i.e., our studies do not include physiological mechanisms.

Conclusion

Cognitive effort remains an elusive concept to capture [13]. We did not find that demand avoidance in the DST, cognitive effort discounting in COGED and rational reasoning items measure the same latent construct of cognitive effort. However, both effort discounting and rational reasoning were related to Need for Cognition and working memory capacity. Demand avoidance in the DST had no association with Need for Cognition or any of the other measures. As both DST and COGED are used frequently as measures of cognitive effort including clinical samples, our findings have large implications for interpretations of previous findings.

If the two tasks are measuring different constructs, then findings with one task should not be interpreted as applying to the other task. Lastly, our work highlights the need for developing new behavioral paradigms for measuring cognitive effort [13]. We recommend considering multiple tasks for estimating the latent construct of sensitivity to cognitive effort costs as well as a rating of perceived mental effort.

Supporting information

S1 File. Details to the experiments, alternative analysis, figures and reliability.
(DOCX)

Acknowledgments

We thank Wouter Kool for providing the Matlab script of the demand selection task. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions

Conceptualization: Martin Jensen Mækelæ, Andrew Westbrook, Gerit Pfuhl.

Data curation: Kristoffer Klevjer, Andrew Westbrook, Gerit Pfuhl.

Formal analysis: Andrew Westbrook, Gerit Pfuhl.

Funding acquisition: Andrew Westbrook, Gerit Pfuhl.

Investigation: Martin Jensen Mækelæ, Kristoffer Klevjer, Andrew Westbrook, Noah S. Eby, Rikke Eriksen, Gerit Pfuhl.

Methodology: Martin Jensen Mækelæ, Kristoffer Klevjer, Andrew Westbrook, Noah S. Eby, Rikke Eriksen, Gerit Pfuhl.

Project administration: Andrew Westbrook, Gerit Pfuhl.

Resources: Andrew Westbrook, Gerit Pfuhl.

Software: Gerit Pfuhl.

Supervision: Andrew Westbrook, Gerit Pfuhl.

Validation: Gerit Pfuhl.

Visualization: Kristoffer Klevjer, Rikke Eriksen, Gerit Pfuhl.

Writing – original draft: Martin Jensen Mækelæ, Gerit Pfuhl.

Writing – review & editing: Martin Jensen Mækelæ, Kristoffer Klevjer, Andrew Westbrook, Noah S. Eby, Rikke Eriksen, Gerit Pfuhl.

References

1. Hull CL. Principles of behavior: an introduction to behavior theory. Oxford, England: Appleton-Century; 1943. x, 422–x, p.
2. Allport GW. The nature of prejudice. Oxford, England: Addison-Wesley; 1954.
3. Shenhav A, Botvinick MM, Cohen JD. The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron. 2013; 79(2):217–40. https://doi.org/10.1016/j.neuron.2013.07.007 PMID: 23889930

4. Zénon A, Solopchuk O, Pezzulo G. An information-theoretic perspective on the costs of cognition. Neuropsychologia. 2019; 123:5–18. https://doi.org/10.1016/j.neuropsychologia.2018.09.013 PMID: 30268880

5. Musslick S, Cohen JD, Shenhav A. Estimating the costs of cognitive control from task performance: theoretical validation and potential pitfalls. Madison, WI.: Proceedings of the 40th Annual Meeting of the Cognitive Science Society; 2018. p. 800–5.

6. Sayalı C, Badre D. Neural systems of cognitive demand avoidance. Neuropsychologia. 2019; 123:41–54. https://doi.org/10.1016/j.neuropsychologia.2018.06.016 PMID: 29944865

7. Westbrook A, Lamichhane B, Braver T. The Subjective Value of Cognitive Effort is Encoded by a Domain-General Valuation Network. The Journal of Neuroscience. 2019:3071–18. https://doi.org/10.1523/JNEUROSCI.3071-18.2019 PMID: 30850512

8. Botvinick MM, Huffstetler S, McGuire JT. Effort discounting in human nucleus accumbens. Cogn Affect Behav Neurosci. 2009; 9(1):16–27. https://doi.org/10.3758/CABN.9.1.16 PMID: 19246324

9. Aston-Jones G, Cohen JD. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci. 2005; 28:403–50. https://doi.org/10.1146/annurev.neuro.28.061604.135709 PMID: 16022602

10. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience. 2005; 8(12):1704–11. https://doi.org/10.1038/nn1560 PMID: 16286932

11. Kool W, Botvinick M. A labor/leisure tradeoff in cognitive control. J Exp Psychol Gen. 2014; 143(1):131–41. https://doi.org/10.1037/a0031048 PMID: 23230991

12. Kurzban R, Duckworth A, Kable JW, Myers J. An opportunity cost model of subjective effort and task performance. Behav Brain Sci. 2013; 36(6):661–79. https://doi.org/10.1017/S0140525X12003196 PMID: 24304775

13. Thomson KS, Oppenheimer DM. The "Effort Elephant" in the Room: What Is Effort, Anyway? Perspect Psychol Sci. 2022; 17(6):1633–52. https://doi.org/10.1177/17456916211064896 PMID: 35767344

14. Baumeister RF, Bratslavsky E, Muraven M, Tice DM. Ego depletion: Is the active self a limited resource? Journal of Personality and Social Psychology. 1998; 74(5):1252–65. https://doi.org/10.1037//0022-3514.74.5.1252 PMID: 9599441

15. Bijleveld E, Custers R, Aarts H. The unconscious eye opener: pupil dilation reveals strategic recruitment of resources upon presentation of subliminal reward cues. Psychol Sci. 2009; 20(11):1313–5. https://doi.org/10.1111/j.1467-9280.2009.02443.x PMID: 19788532

16. Kahneman D. Attention and Effort: Englewood Cliffs, NJ: Prentice-Hall; 1973.

17. Braver TS. The variable nature of cognitive control: a dual mechanisms framework. Trends in Cognitive Sciences. 2012; 16(2):106–13. https://doi.org/10.1016/j.tics.2011.12.010 PMID: 22245618

18. Feng SF, Schwemmer M, Gershman SJ, Cohen JD. Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. Cogn Affect Behav Neurosci. 2014; 14(1):129–46. https://doi.org/10.3758/s13415-013-0236-9 PMID: 24481850

19. Holroyd CB. The waste disposal problem of effortful control. Motivation and cognitive control. Frontiers of cognitive psychology. New York, NY, US: Routledge/Taylor & Francis Group; 2016. p. 235–60.

20. Camerer CF, Hogarth RM. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. Journal of Risk and Uncertainty. 1999; 19(1):7–42.

21. Bonner SE, Sprinkle GB. The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. Accounting, Organizations and Society. 2002; 27(4):303–45.

22. Engelmann JB, Damaraju E, Padmala S, Pessoa L. Combined effects of attention and motivation on visual task performance: transient and sustained motivational effects. Front Hum Neurosci. 2009; 3:4. https://doi.org/10.3389/neuro.09.004.2009 PMID: 19434242

23. Kanwal J, Smith K, Culbertson J, Kirby S. Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. Cognition. 2017; 165:45–52. https://doi.org/10.1016/j.cognition.2017.05.001 PMID: 28494263

24. Shah AK, Oppenheimer DM. Heuristics made easy: An effort-reduction framework. Psychological Bulletin. 2008; 134(2):207–22. https://doi.org/10.1037/0033-2909.134.2.207 PMID: 18298269

25. Cacioppo JT, Petty RE. The need for cognition. Journal of Personality and Social Psychology. 1982; 42(1):116–31.

26. Hussey I, Hughes S. Hidden Invalidity Among 15 Commonly Used Measures in Social and Personality Psychology. Advances in Methods and Practices in Psychological Science. 2020; 3(2):166–84.

27. Paulhus DL, Vazire S. The self-report method. Handbook of research methods in personality psychology. New York, NY, US: The Guilford Press; 2007. p. 224–39.

28. Reddy LF, Horan WP, Barch DM, Buchanan RW, Dunayevich E, Gold JM, et al. Effort-Based Decision-Making Paradigms for Clinical Trials in Schizophrenia: Part 1-Psychometric Characteristics of 5 Paradigms. Schizophr Bull. 2015; 41(5):1045–54.

29. Horan WP, Reddy LF, Barch DM, Buchanan RW, Dunayevich E, Gold JM, et al. Effort-Based Decision-Making Paradigms for Clinical Trials in Schizophrenia: Part 2-External Validity and Correlates. Schizophr Bull. 2015; 41(5):1055–65.

30. Kool W, McGuire JT, Rosen ZB, Botvinick MM. Decision making and the avoidance of cognitive demand. J Exp Psychol Gen. 2010; 139(4):665–82. https://doi.org/10.1037/a0020198 PMID: 20853993

31. Pfuhl G. Two strings to choose from: do ravens pull the easier one? Animal cognition. 2012; 15(4):549–57. https://doi.org/10.1007/s10071-012-0483-0 PMID: 22437450

32. Westbrook A, Kester D, Braver TS. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. PloS one. 2013; 8(7):e68210. https://doi.org/10.1371/journal.pone.0068210 PMID: 23894295

33. ten Velden Hegelstad W, Kreis I, Tjelmeland H, Pfuhl G. Psychosis and psychotic-like symptoms affect cognitive abilities but not motivation in a foraging task. Frontiers in Psychology. 2020;11.

34. Gatzke-Kopp LM, Ram N, Lydon-Staley DM, DuPuis D. Children's Sensitivity to Cost and Reward in Decision Making across Distinct Domains of Probability, Effort, and Delay. Journal of Behavioral Decision Making. 2018; 31(1):12–24. https://doi.org/10.1002/bdm.2038 PMID: 29353962

35. Toplak ME, West RF, Stanovich KE. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking & Reasoning. 2014; 20(2):147–68.

36. West RF, Toplak ME, Stanovich KE. Heuristics and Biases as Measures of Critical Thinking: Associations with Cognitive Ability and Thinking Dispositions. Journal of Educational Psychology. 2008; 100 (4):930–41.

37. Toplak ME, West RF, Stanovich KE. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory & Cognition. 2011; 39(7):1275. https://doi.org/10.3758/s13421-011-0104-1 PMID: 21541821

38. Pennycook G, Fugelsang JA, Koehler DJ. What makes us think? A three-stage dual-process model of analytic engagement. Cognitive psychology. 2015; 80:34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001 PMID: 26091582

39. Chang WC, Westbrook A, Strauss GP, Chu AOK, Chong CSY, Siu CMW, et al. Abnormal cognitive effort allocation and its association with amotivation in first-episode psychosis. Psychological Medicine. 2020; 50(15):2599–609. https://doi.org/10.1017/S0033291719002769 PMID: 31576787

40. Culbreth AJ, Gold JM, Cools R, Barch DM. Impaired Activation in Cognitive Control Regions Predicts Reversal Learning in Schizophrenia. Schizophrenia Bulletin. 2016; 42(2):484–93. https://doi.org/10.1093/schbul/sbv075 PMID: 26049083

41. Gold JM, Kool W, Botvinick MM, Hubzin L, August S, Waltz JA. Cognitive effort avoidance and detection in people with schizophrenia. Cogn Affect Behav Neurosci. 2015; 15(1):145–54. https://doi.org/10.3758/s13415-014-0308-5 PMID: 24957405

42. Puveendrakumaran P, Fervaha G, Caravaggio F, Remington G. Assessing analytic and intuitive reasoning using the cognitive reflection test in young patients with schizophrenia. Psychiatry Research. 2020; 284:112683. https://doi.org/10.1016/j.psychres.2019.112683 PMID: 31818543

43. Strobel A, Wieder G, Paulus PC, Ott F, Pannasch S, Kiebel SJ, et al. Dispositional cognitive effort investment and behavioral demand avoidance: Are they related? PloS one. 2020; 15(10):e0239817. https://doi.org/10.1371/journal.pone.0239817 PMID: 33052978

44. Stanovich KE. What intelligence tests miss. The psychology of rational thought: Yale University Press; 2009.

45. Trippas D, Pennycook G, Verde MF, Handley SJ. Better but still biased: Analytic cognitive style and belief bias. Thinking & Reasoning. 2015; 21(4):431–45.

46. Shenhav A, Rand DG, Greene JD. The relationship between intertemporal choice and following the path of least resistance across choices, preferences, and beliefs. Judgment and Decision Making. 2017; 12(1):1–18.

47. Stanovich KE, West RF, Toplak ME. The rationality quotient. Toward a test of rational thinking. Cambridge, Massachusetts: MIT Press; 2016.

48. Frederick S. Cognitive Reflection and Decision Making. Journal of Economic Perspectives. 2005; 19 (4):25–42.

49. Raoelison M, Boissin E, Borst G, De Neys W. From slow to fast logic: the development of logical intuitions. Thinking & Reasoning. 2021; 27(4):599–622.

50. Stupple EJN, Pitchford M, Ball LJ, Hunt TE, Steel R. Slower is not always better: Response-time evidence clarifies the limited role of miserly information processing in the Cognitive Reflection Test. PloS one. 2017; 12(11):e0186404. https://doi.org/10.1371/journal.pone.0186404 PMID: 29099840

51. Thompson VA, Pennycook G, Trippas D, Evans J. Do smart people have better intuitions? J Exp Psychol Gen. 2018; 147(7):945–61. https://doi.org/10.1037/xge0000457 PMID: 29975089

52. Mækelæ MJ, Pfuhl G. Deliberate reasoning is not affected by language. PloS one. 2019; 14(1): e0211428. https://doi.org/10.1371/journal.pone.0211428 PMID: 30703137

53. Raoelison M, Thompson VA, De Neys W. The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. Cognition. 2020; 204:104381.

54. Mækelæ MJ, Moritz S, Pfuhl G. Are psychotic experiences related to poorer reflective reasoning? Frontiers in psychology. 2018; 9:122. https://doi.org/10.3389/fpsyg.2018.00122 PMID: 29483886

55. Attali Y, Bar-Hillel M. The false allure of fast lures. Judgment and Decision Making. 2020; 15:93–111.

56. Gong T, Young AG, Shtulman A. The Development of Cognitive Reflection in China. Cognitive Science. 2021; 45(4):e12966. https://doi.org/10.1111/cogs.12966 PMID: 33873237

57. Reddy LF, Reavis EA, Wynn JK, Green MF. Pupillary responses to a cognitive effort task in schizophrenia. Schizophr Res. 2018; 199:53–7. https://doi.org/10.1016/j.schres.2018.03.005 PMID: 29526458

58. Cacioppo JT, Petty RE, Feinstein JA, Jarvis WBG. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. Psychological Bulletin. 1996; 119(2):197–253.

59. Inzlicht M, Shenhav A, Olivola CY. The Effort Paradox: Effort Is Both Costly and Valued. Trends Cogn Sci. 2018; 22(4):337–49. https://doi.org/10.1016/j.tics.2018.01.007 PMID: 29477776

60. Thomson KS, Oppenheimer DM. Investigating an alternate form of the cognitive reflection test. Judgment and Decision Making. 2016; 11(1):99–113.

61. Patzelt EH, Kool W, Millner AJ, Gershman SJ. The transdiagnostic structure of mental effort avoidance. Scientific Reports. 2019; 9(1):1689. https://doi.org/10.1038/s41598-018-37802-1 PMID: 30737422

62. Culbreth A, Westbrook A, Barch D. Negative symptoms are associated with an increased subjective cost of cognitive effort. J Abnorm Psychol. 2016; 125(4):528–36. https://doi.org/10.1037/abn0000153 PMID: 26999282

63. Westbrook A, van den Bosch R, Määttä JI, Hofmans L, Papadopetraki D, Cools R, et al. Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. Science. 2020; 367 (6484):1362–6. https://doi.org/10.1126/science.aaz5891 PMID: 32193325

64. Chevalier N. Willing to Think Hard? The Subjective Value of Cognitive Effort in Children. Child Development. 2018; 89(4):1283–95.

65. Stagnaro MN, Pennycook G, Rand DG. Performance on the Cognitive Reflection Test is stable across time. Judgment and Decision Making. 2018; 13(3):260–7.

66. Kreis I, Moritz S, Pfuhl G. Objective versus subjective effort in schizophrenia. Frontiers in Psychology. 2020; 11:1469. https://doi.org/10.3389/fpsyg.2020.01469 PMID: 32742265

67. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Hancock PA, Meshkati N, editors. Advances in Psychology. 52: North-Holland; 1988. p. 139–83.

68. Owen AM, McMillan KM, Laird AR, Bullmore E. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Hum Brain Mapp. 2005; 25(1):46–59. https://doi.org/10.1002/hbm.20131 PMID: 15846822

69. Kleiner M, Brainard DH, Pelli DG. What's new in Psychtoolbox-3. Perception 36 ECVP Abstract Supplement. 2007.

70. Pelli DG. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis. 1997; 10(4):437–42. PMID: 9176953

71. Brainard DH. The Psychophysics Toolbox. Spatial Vision. 1997; 10(4):433–6. PMID: 9176952

72. Levesque HJ. Knowledge Representation and Reasoning. Annual Review of Computer Science. 1986; 1(1):255–87.

73. Koehler D J., James G. Probability matching and strategy availability. Memory & Cognition. 2010; 38 (6):667–76. https://doi.org/10.3758/MC.38.6.667 PMID: 20852231

74. Teigen KH, Keren G. Waiting for the bus: when base-rates refuse to be neglected. Cognition. 2007; 103 (3):337–57. https://doi.org/10.1016/j.cognition.2006.03.007 PMID: 16723123

75. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping Doctors and Patients Make Sense of Health Statistics. Psychol Sci Public Interest. 2007; 8(2):53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x PMID: 26161749

76. Toplak ME, Liu E, Macpherson R, Toneatto T, Stanovich KE. The reasoning skills and thinking dispositions of problem gamblers: A dual-process taxonomy. Journal of Behavioral Decision Making. 2007; 20 (2):103–24.

77. Stanovich KE, West RF. Individual differences in rational thought. Journal of Experimental Psychology: General. 1998; 127(2):161–88.

78. Smullyan RM. What is the name of this book? The riddle of Dracula and other logical puzzles. Englewood Cliffs, NJ: Prentice-Hall; 1978.

79. Lehman DR, Lempert RO, Nisbett RE. The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. American Psychologist. 1988; 43(6):431–42.

80. Cacioppo JT, Petty RE, Feng Kao C. The Efficient Assessment of Need for Cognition. Journal of Personality Assessment. 1984; 48(3):306–7. https://doi.org/10.1207/s15327752jpa4803_13 PMID: 16367530

81. Hart SG. Nasa-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2006; 50(9):904–8.

82. Treadway MT, Buckholtz JW, Schwartzman AN, Lambert WE, Zald DH. Worth the 'EEfRT'? The effort expenditure for rewards task as an objective measure of motivation and anhedonia. PloS one. 2009; 4 (8):e6598. https://doi.org/10.1371/journal.pone.0006598 PMID: 19672310

83. Xu X, Demos KE, Leahey TM, Hart CN, Trautvetter J, Coward P, et al. Failure to Replicate Depletion of Self-Control. PloS one. 2014; 9(10):e109950. https://doi.org/10.1371/journal.pone.0109950 PMID: 25333564

84. Team R. RStudio: Integrated Development for R. RStudio, PBC, Bosten, MA; 2022.

85. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software 2010; 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03

86. Eldar E, Felso V, Cohen JD, Niv Y. A pupillary index of susceptibility to decision biases. Nat Hum Behav. 2021; 5(5):653–62. https://doi.org/10.1038/s41562-020-01006-3 PMID: 33398147

87. Erceg N, Galić Z, Bubić A, Jelić D. Who detects and why: how do individual differences in cognitive characteristics underpin different types of responses to reasoning tasks? Thinking & Reasoning. 2022:1–49.

88. Otero I, Salgado JF, Moscoso S. Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. Intelligence. 2022; 90:101614.

89. Stanovich KE. Miserliness in human cognition: the interaction of detection, override and mindware. Thinking & Reasoning. 2018; 24(4):423–44.

90. Pennycook G, Rand DG. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition. 2019; 188:39–50. https://doi.org/10.1016/j.cognition.2018.06.011 PMID: 29935897

91. Zerna J, Scheffel C., Kührt C., & Strobel A. When easy is not preferred: An effort discounting paradigm for estimating subjective values of tasks. Nature Scientific Reports. 2022;Stage 1 Registered Report.

92. Sandra DA, Otto AR. Cognitive capacity limitations and Need for Cognition differentially predict reward-induced cognitive effort expenditure. Cognition. 2018; 172:101–6. https://doi.org/10.1016/j.cognition.2017.12.004 PMID: 29247878

93. Kramer A-W, Van Duijvenvoorde ACK, Krabbendam L, Huizenga HM. Individual differences in adolescents' willingness to invest cognitive effort: Relation to need for cognition, motivation and cognitive capacity. Cognitive Development. 2021; 57:100978.

94. Blaise M, Marksteiner T, Krispenz A, Bertrams A. Measuring Motivation for Cognitive Effort as State. Frontiers in Psychology. 2021;12. https://doi.org/10.3389/fpsyg.2021.785094 PMID: 34956008

95. Kashdan TB, Stiksma MC, Disabato DJ, McKnight PE, Bekier J, Kaji J, et al. The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. Journal of Research in Personality. 2018; 73:130–49.

96. Lopez-Gamundi P, Wardle MC. The cognitive effort expenditure for rewards task (C-EEfRT): A novel measure of willingness to expend cognitive effort. Psychological assessment. 2018; 30(9):1237–48. https://doi.org/10.1037/pas0000563 PMID: 29620381

Supplementary material for: “Is it cognitive effort you measure? Comparing three task paradigms to the Need for Cognition scale”

Additional information for Study 1

Results: Participants had a demand avoidance (preference for the high demand deck: M = .42, SD = .17) that was not significantly different from indifference between the high and low demand option (Z = 0.674, p < 0.5, d = .077). Accuracy in the DST was high (M = 0.94, SD = 0.07). The COGED paradigm showed significantly larger monetary discounting with increasing load levels F(4, 380) = 24.109, p < .001, η^2 = .202. The average indifference point was 1.46 (SD = .44). Across all six n-back levels the discriminability d’ was 2.06 (SD = .7) and perceived mental effort was 66 (SD = 18.2) on a scale from 0 to 100. The Need for Cognition Score ranged from 35 to 86, with a mean of 65.1 (SD = 11.3). The average O-span performance was 51.82 (SD = 16.81).

None of the tasks was related to working memory as assessed with the O-Span task, all τ < .01, all p > .2. O-Span and d’ were not associated, τ = .101, p = .2.

Additional information for Study 2

Results: Like study 1, participants performed well on the O-span (M = 50.78, SD = 18.57), their NCS ranged from 36 to 84 with a mean of 66.56 (SD = 9.82). They had on average over half of the rational reasoning items correct (M = 10.04, SD = 4.04). The COGED paradigm showed significantly larger monetary discounting with increasing load levels, F(4, 240) = 24.734, p < .001, η^2 = .191. The average indifference point was 1.58 (SD = .42). Across all six n-back levels the discriminability d’ was 1.68 (SD = .76).

O-Span was positively correlated with rational reasoning, τ = .209, p = .008 and d’, τ = .153, p = .045, but not with cognitive effort discounting, τ = .096, p = .226; or Need for Cognition, τ = .006, p = .422.

Additional information for Study 3

Methods: Regarding the rational reasoning items we used items 2-7 from the Cognitive Reflection Test [1], one fully disjunctive reasoning problem “the marriage problem” [2], one probability matching task [3], one probability estimation task “the bus problem” [4], one making sense of medical results problem [5], one Bayesian reasoning problem [6], one covariation detection problem [7], one knight and knave problem [8], one conditional reasoning problem [9].

Results: The Need for Cognition score ranged from 42 to 81 with a mean of 61.73 (SD = 9.15). The indifference point was M = 1.18 (SD = .56), discriminability d' was M = 2.08 (SD = .66) and perceived mental demand was high, M = 73.13 (SD = 16.04). The rational reasoning score was M = 5.08 (SD = 2.49) and perceived mental demand was M = 67.2 (SD = 14.95).

Additional information for Study 4

Methods: Filler tasks in study 4: On day 1 participants did also the Bullshit receptivity scale [10]), the Effort expenditure for rewards task (EEfRT, [11]) and N-TLX$_{EEfRT}$. On Day 2 they did after the DST and NCS a Handgrip task [12]

Results: For the Demand Selection Task, accuracy was high on Day 1 (M = 0.98, SD = 0.02) and Day 2 (M = 0.96, SD = 0.08). The median of high demand choice was 0.49 for Day 1 and Day 2. Demand avoidance was not different from .5, neither on day Day 1 (Z = .21, p = .834, d = .034) nor on Day 2 (Z = .238, p = .812, d = .039). Debriefing identified 5 participants on Day 1 that identified the manipulation (two types of decks) and 11 participants who might have. On Day 2 the manipulation was found by 13 participants, and another 22 might have found it. Among those who discovered the manipulation the high demand avoidance was .6 on Day 1 and .54 on Day 2. One participant who noticed the manipulation said they tried not to exploit it. On Day 2 demand avoidance was not related to NCS (τ = .165, p = .16).

The COGED paradigm showed significant increases in monetary discounting with increasing load levels, F(2, 114) = 4.432, p = .014, η^2 = .072. The average indifference point was 1.16 (SD = .51) and discriminability d' had M = 2.37 (SD = .3). Participants had on average half of the rational reasoning items correct, M = 7.0 (SD = 2.7). The Need for Cognition score had M = 62 (SD = 10.11) on day 1 and M = 65 (SD = 11.5) on day 2. Need for Cognition had good internal consistency on Day 1 (Cronbach's α = .83) and Day 2 (Cronbach's α = .89), and good reliability across the two testing sessions (r = .823, p < .001).

Additional information for Study 5

The Need for Cognition score ranged from 33 to 83, M = 63.46 (SD = 9.77). On average participants solved half of the items correctly, M = 6.82 (SD = 2.61) and perceived mental demand was high, M = 79.11 (SD = 18.87). Accuracy on the demand selection task was high, M = 0.97 (SD = 0.04), perceived mental demand low, M = 45.56 (SD = 24.5). Participants avoided cognitive demand, M = .45 (SD = .12, Md = 0.47) but demand preference was not significantly different from chance (Z test, Z = .37, p < 0.711, d = .055).

Additional information for Study 6

Methods: Regarding the rational reasoning items we used items 2-7 from the Cognitive Reflection Test [1], one fully disjunctive reasoning problem, “the marriage problem” [2], one knight and knave problem [8], one conditional reasoning problem [9], one covariation problem [13], one base rate problem [14], one making sense of medical results problem [5].

Results: The two samples had similar NCS scores, t(197.38) = .838, p = .403, M_{pooled} = 61.59, SD = 10.94, similar working memory capacity, t(203.32) = -.12, p = .904, M_{pooled} = 3.31, SD = 2.13, similar accuracy in the DST, t(208.31) = -.062, p = .95, M_{pooled} = .92, SD = .076 and similar demand avoidance, t(127.77) = 1.074, p = .285, M_{pooled} = .44, SD = .23. The rational reasoning score was higher in the UiT sample (M = 6.01, SD = 2.71) compared to the Prolific sample (M = 4.83, SD = 2.62). This difference was significant; t(163.75) = -3.633, p = .00037). The average indifference point in COGED was higher in the Prolific sample (M = 1.24, SD = .57) than in the UiT sample (M = 1.01, SD = .58). This difference was significant, t(164.97) = 3.157, p = .002).

Differences between the six studies

We expressed the rational reasoning score as percentage correct. A one-way ANOVA yielded a significant difference between studies (F(4, 577) = 11.253, p < .001, η^2 = .072). Post-hoc Tukey HSD found a significant difference between study 2 and 3 (t = 6.389, p < .001), study 2 and 6 (t = 5.108, p < .001), study 3 and 4 (t = 2.741, p = .049), study 3 and 5 (t = 3.271, p = .01) and study 3 and 6 (t = 2.786, p = .044). Since the total number of items differed across studies, we z-scored the values.

One-way ANOVA yielded a significant difference between studies for cognitive effort discounting (indifference point) in the COGED (F(4, 585) = 12.683, p < .001, η^2 = .08. Post-hoc Tukey HSD test was significant for study 1 versus study 3 (t = 3.571, p = .004) and study 1 versus study 6 (t = 4.372, p < .001), study 2 differed from study 3 (t = 4.869, p < .001), study 4 (t = 2.817, p = .04) and study 6 (t = 6.05, p < .001). Since study 1 and 2 used up to 6-back whereas studies 3-6 used only up to 4-back, we z-scored the indifference point values.

One-way ANOVA yielded no significant difference between studies for the DST (F(3, 472) = .386, p = .763, η^2 = .002). The DST was significantly below .5 (M = .44, SD = .2), i.e., on average participants avoided the high demand option (t(475) = 6.555, p < .001, Cohen’s d= .3).

One-way ANOVA yielded a significant difference between studies for NCS (F(5, 634) = 4.058, p < .001, η^2 = .031). Post-hoc Tukey HSD test was significant for the comparison of study 2 with study 3 (t = 3.084, p = .026) and study 2 with study 6 (t = 3.96, p = .001). No other comparison was significant (see SOM for details).

One-way ANOVA yielded a significant difference between studies for d’ (F(4, 588) = 19.281, p < .001, η^2 = .116). Post-hoc Tukey HSD test was significant for the comparison of study 1 with study 6 (t = 4.467, p < .001), study 2 with study 3 (t = 2.849, p = .037), study 2 with study 6 (t = 7.428, p < .001), study 3 with study 6 (t = 3.887, p < .001) and study 4 with study 6 (t = 3.136, p = .015).

Study-wise bi-variate scatterplots

Figure S1: Need for Cognition Score (NCS) and d' (dprime) from n-back phase from COGED

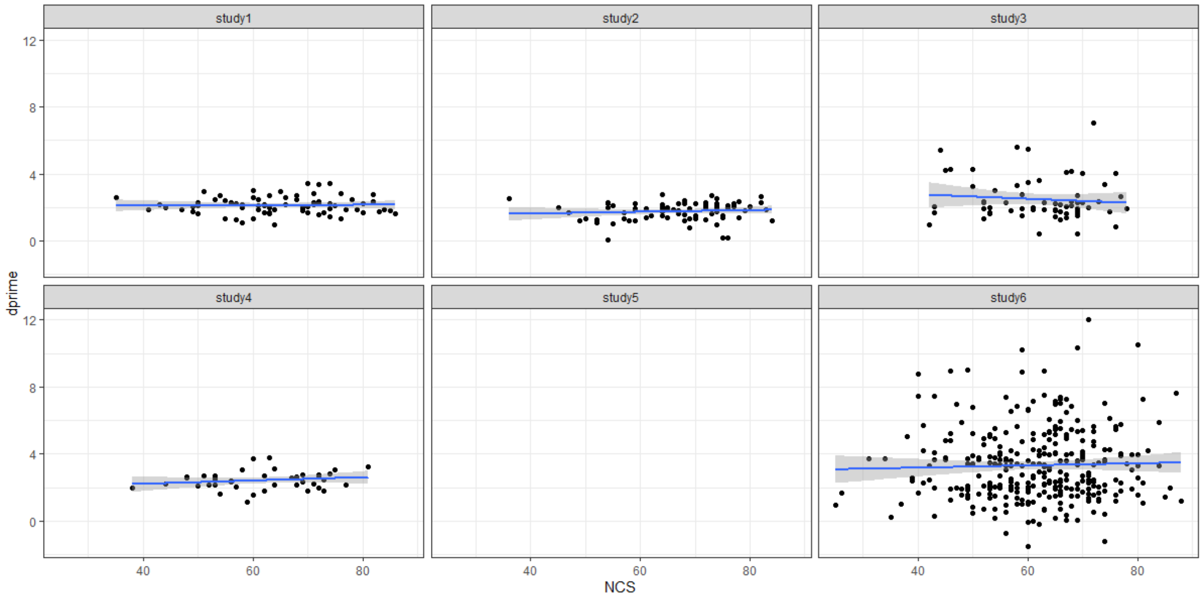


Figure S2: Need for Cognition Score (NCS) and average indifference point from COGED

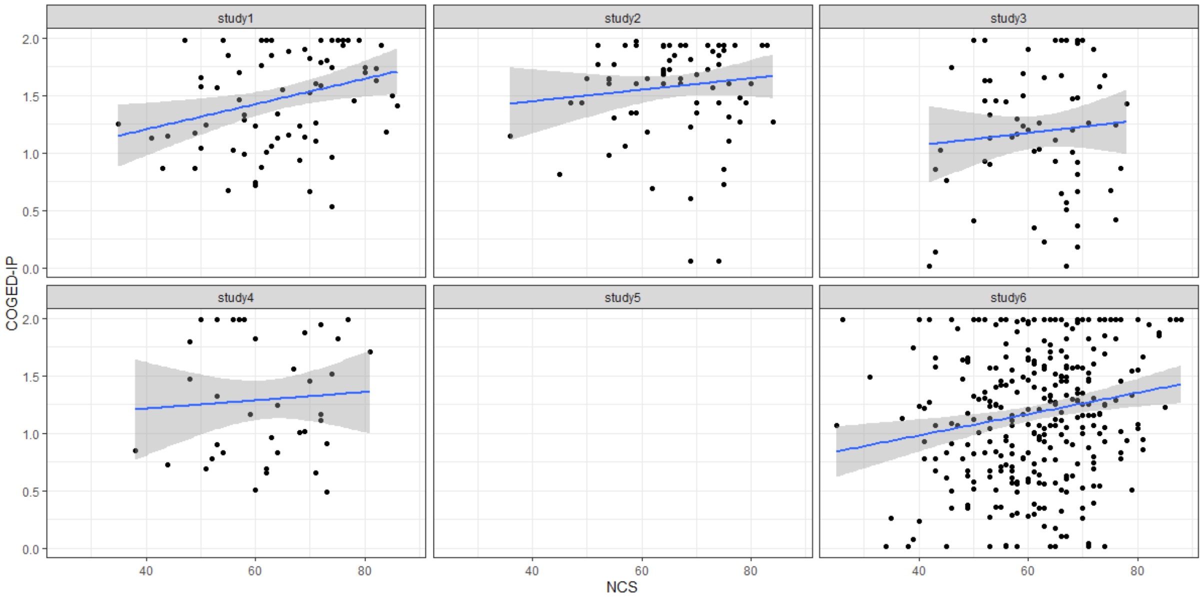


Figure S3: Need for Cognition Score (NCS) and proportions high demand choices in DST

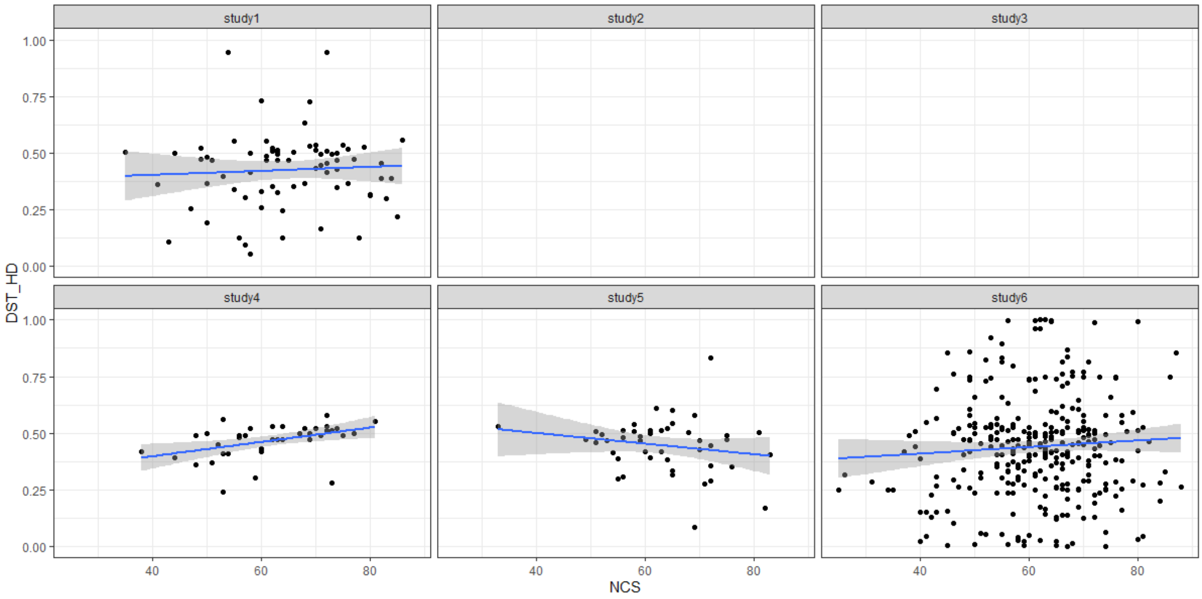


Figure S4: Need for Cognition Score (NCS) and rational reasoning score

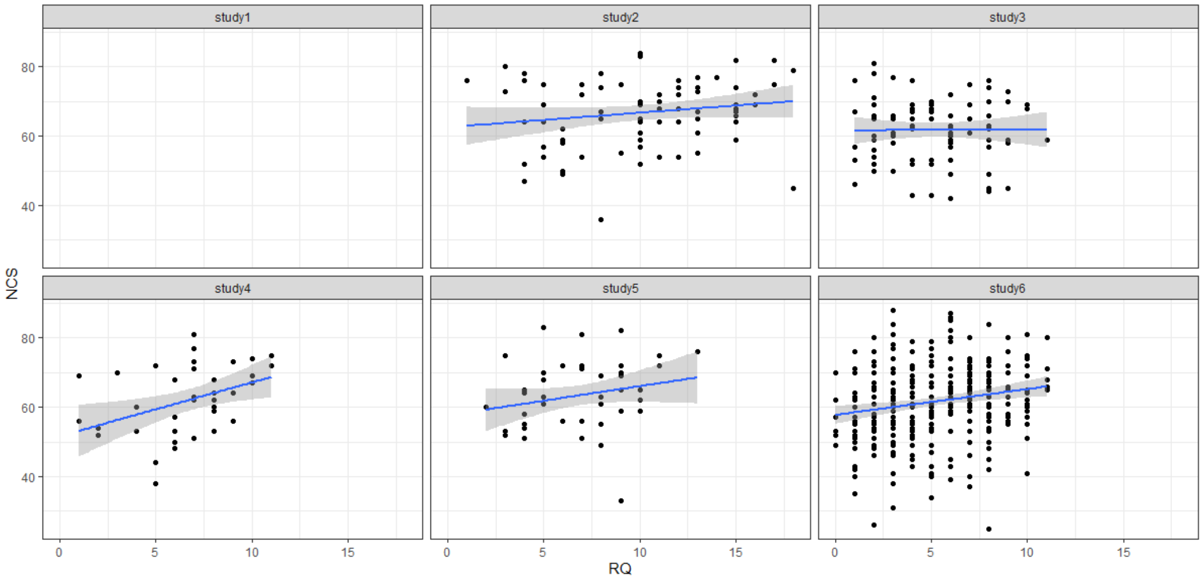


Figure S5 d‘ and indifference point from COGED

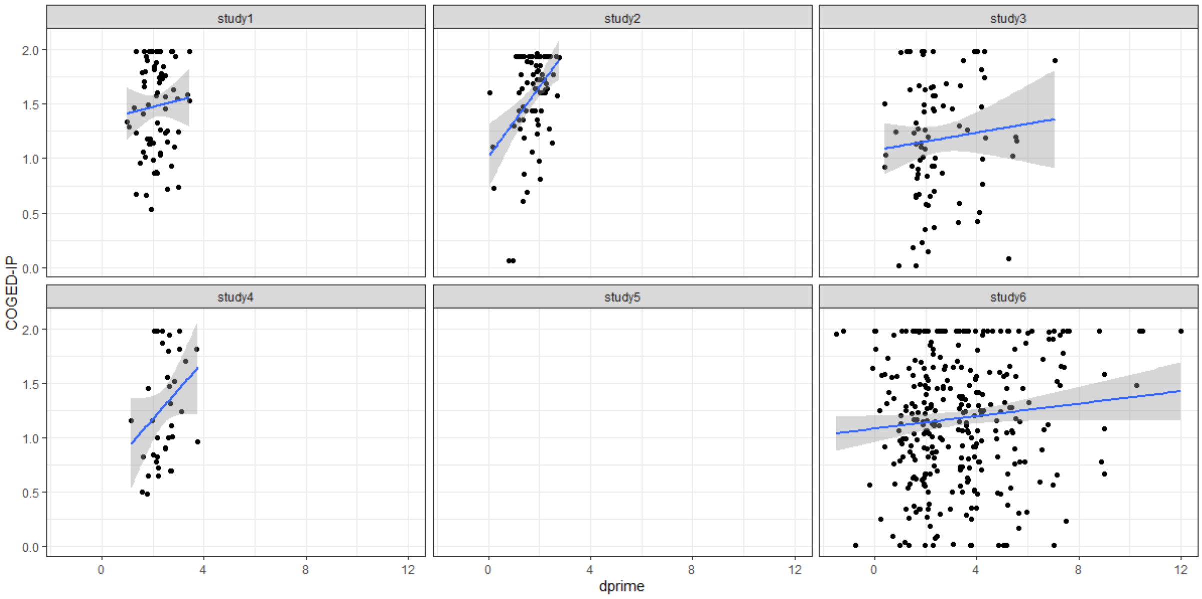


Figure S6 d‘ and proportion high demand choices in DST

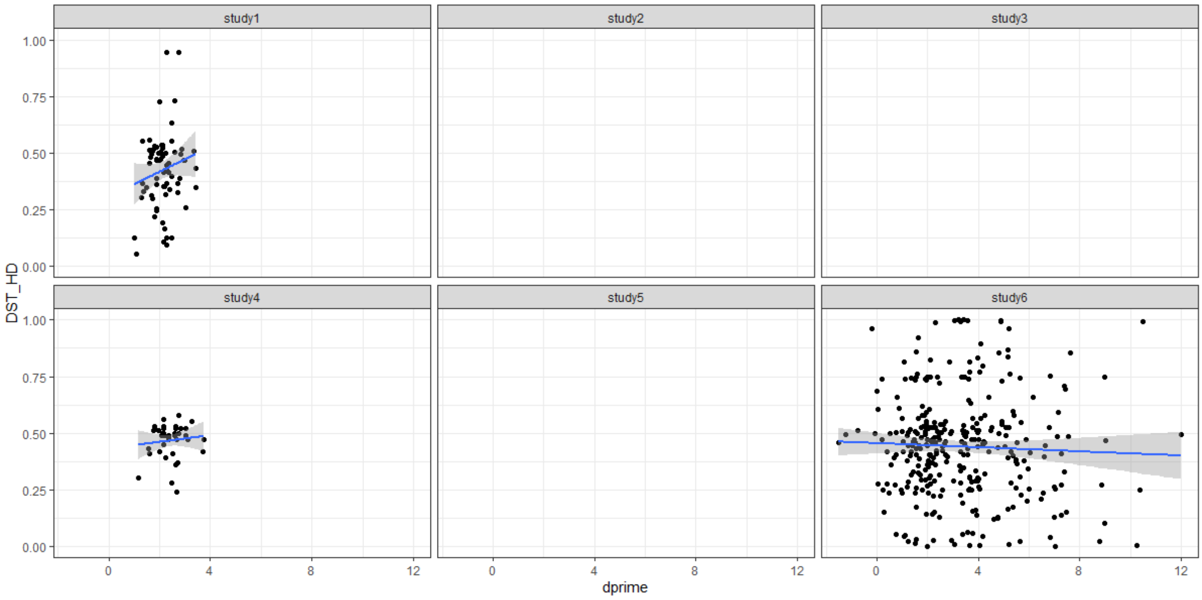


Figure S7 d‘ and rational reasoning score

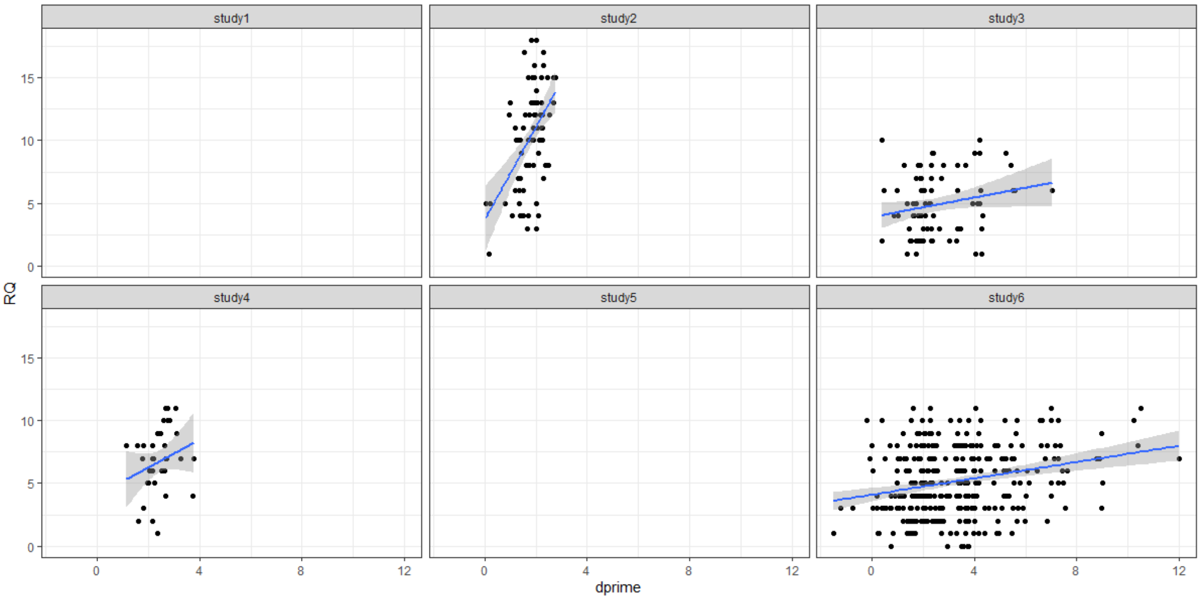


Figure S8 indifference point from COGED and proportion high demand choices in DST

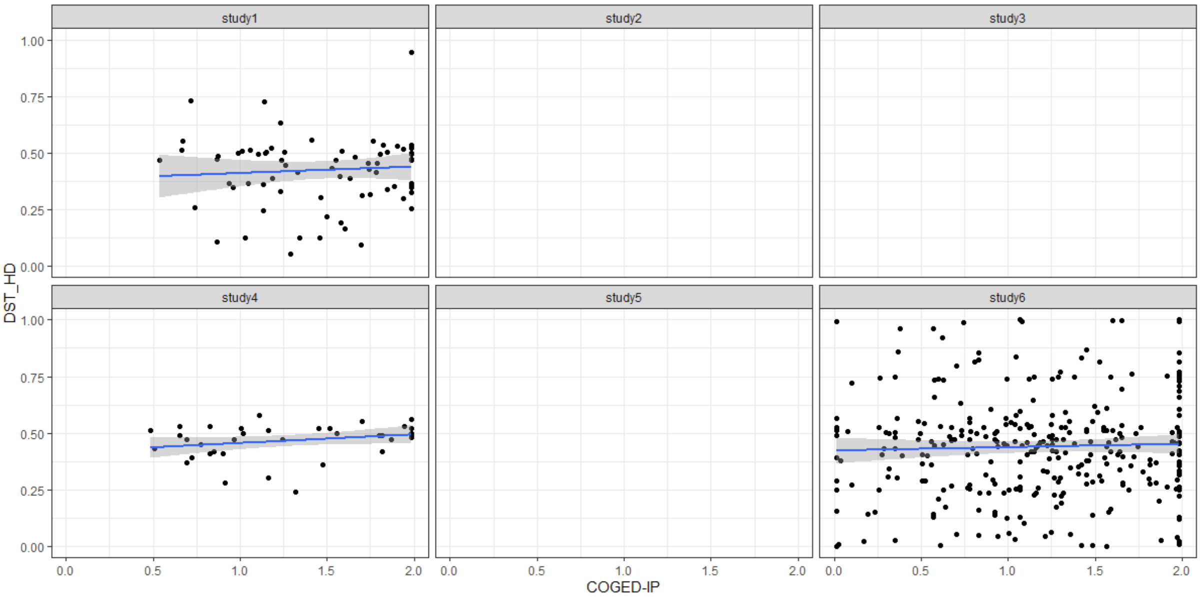


Figure S9 indifference point from COGED and rational reasoning score

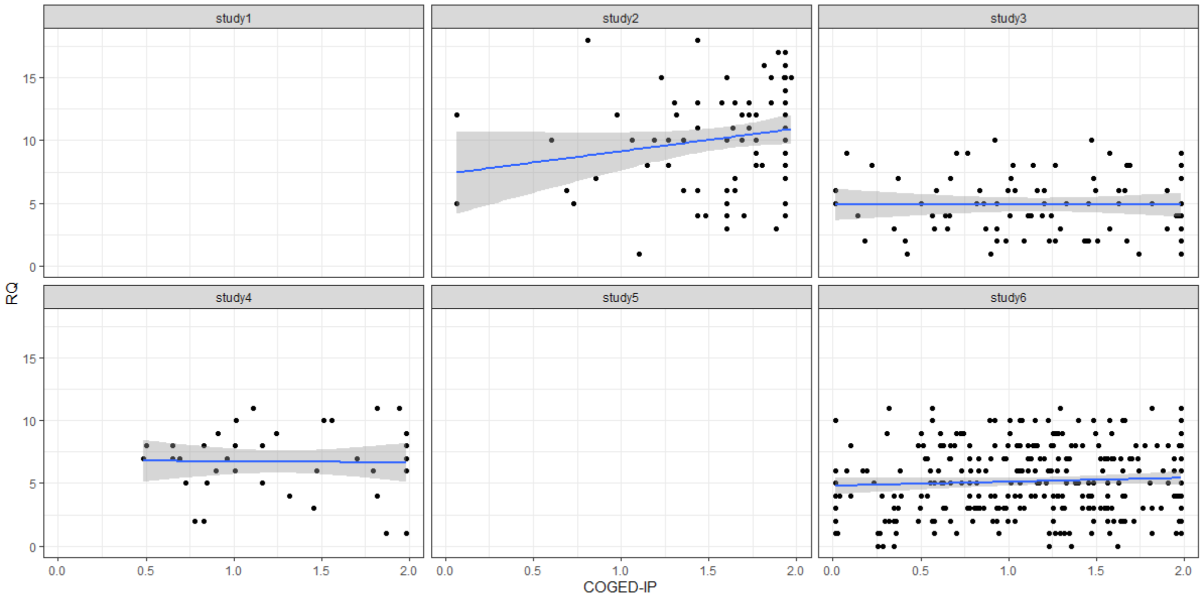
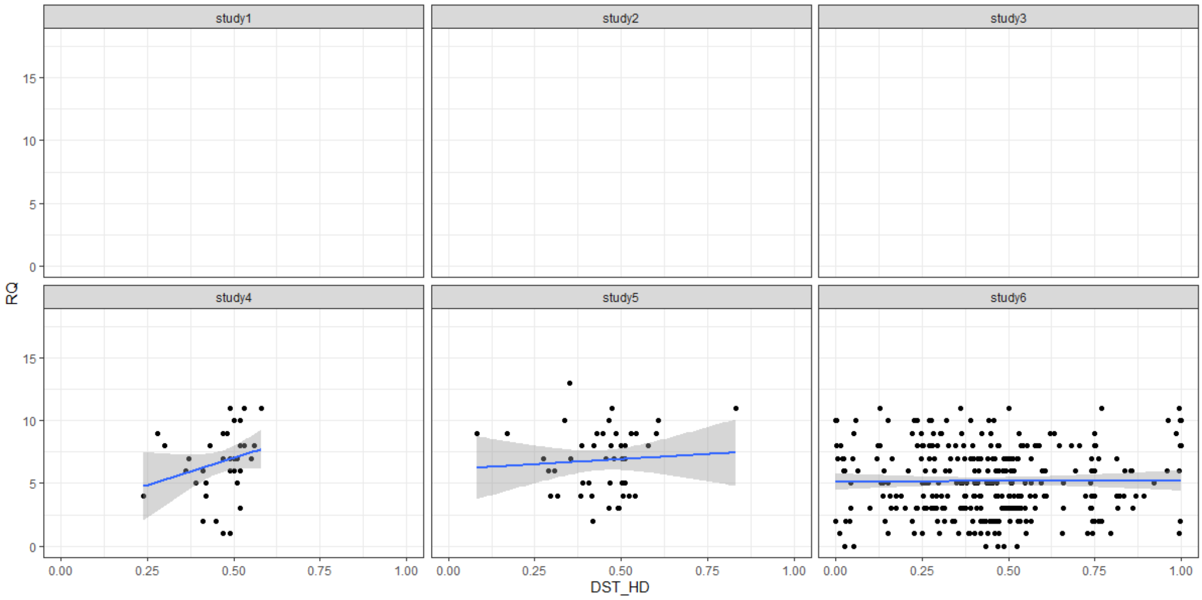


Figure S10 proportion high demand choices in DST and rational reasoning score



Alternative analysis: Pooling after z-scoring

We z-scored all values per study to compare across studies. The Pearson product-moment correlations with p-value and Bayes Factor (BF) are shown in Table 1. BF_{10} < .3 provides support for an absence of a relationship between the tasks. BF_{10} > 3 provides support for the hypothesis that there is an association between the tasks.

Table S1. Pearson's Correlations and Bayes Factors

			n	Pearson's r	p	Lower 95% CI	Upper 95% CI	BF_{10}
Effort discounting, AIP	-	Demand Avoidance	430	0.0669	0.1661	-0.0278	0.1605	**0.1571**
Effort discounting, AIP	-	Rational reasoning score	516	0.0535	0.2254	-0.0330	0.1391	**0.1147**

Table S1. Pearson's Correlations and Bayes Factors

		n	Pearson's r	p	Lower 95% CI	Upper 95% CI	BF$_{10}$
Effort discounting, AIP	- Need for Cognition score	577	0.1769	< .001	0.0967	0.2549	473.0110
Effort discounting, AIP	- n-back d’	591	0.1092	0.0079	0.0288	0.1882	1.7421
Demand Avoidance	- Rational reasoning score	402	0.0326	0.5146	-0.0654	0.1300	**0.0772**
Demand Avoidance	- Need for Cognition score	478	0.0692	0.1310	-0.0206	0.1579	**0.1786**
Demand Avoidance	- n-back d’	431	0.0067	0.8897	-0.0878	0.1011	**0.0609**
Rational reasoning score	- Need for Cognition score	568	0.1768	< .001	0.0960	0.2554	409.2896
Rational reasoning score	- n-back d’	519	0.2736	< .001	0.1921	0.3514	2.7843e +7
Need for Cognition score	- n-back d’	579	0.0420	0.3126	-0.0396	0.1231	**0.0866**

Legend: BF$_{10}$ in bold support the null hypothesis, BF$_{10}$ in italic support the alternative hypothesis. P value in italic support the alternative hypothesis

There are no differences to the meta-analytical approach regarding significance. This analysis complements the analysis reported in the main text by providing Bayes Factors and hence support for the null hypothesis.

Lab vs online studies

The demand selection task yielded similar demand avoidance in lab and online samples, t(456.017) = .049, p = .961, Cohen’s d = .004. The Need for Cognition score was higher among lab- than online participants, t(633.279) = 2.878, p = .004, Cohen’s d = .228. Working memory capacity was higher in the online than the lab samples, t(434.607) = -8.64, p < .001, Cohen’s d = .693. The average indifference point was higher in the lab than online sample, t(587.977) = 4.784, p < .001, Cohen’s d = .393. Rational reasoning score (percentage correct) was similar in the lab and the online sample, t(580) = 1.715, p = .087, Cohen’s d = .143.

Study S1: Test – re-test reliability of rationality items

Methods

Ethics

All methods were performed in accordance with the relevant guidelines and regulations and approved by the Institutional Review Board at the Department of Psychology at UiT – The Arctic University of Norway. Written informed consent was obtained from all participants.

Participants

In total 136 participants were recruited at Prolific (prolific.co). Of the 136, 83 completed both sessions and all questionnaires and had sufficient proficiency in English (assessed with the Word sum test, cut-off was 3 out of 10). 36 of the participants were women (aged 18- 49) and 46 were men (aged 18-64), one indicated as gender other. The experiment was conducted online, allowing people from several countries to participate (24% from Poland, 13.5% Portugal, 12.5% Italy, 10% England, remaining were from 13 other countries). Participants received ca. £10 pounds after completion. Participation was voluntary and participants could withdraw their consent at any moment.

Materials

The experiment was conducted in English and all instructions were in English.

Regarding rational reasoning items we used 14 items from the problem solving and reasoning literature:

- One item from [15].
- One item from [16].
- Items two and three from [17].
- Items 4-6 from [18].
- One item from [2].
- One item from [19].
- One item from [14].
- One item from [20].
- One item from [8]
- One item from the Wason selection task (Wason, 1966; as cited in [21]).
- One item from [22].

Although the degree of difficulty varied in these items, deliberate reasoning was required in all to reach the correct answer. The 14 items were sequentially and randomly presented to all participants. Participants had to provide an answer before proceeding to the next item. There were no time limits for answering the RQ items in session 1.

In session 1 we measured also cognitive abilities with the Berlin numeracy test [23] and the word sum test [24], as well as we used the Need for Cognition scale. These questionnaires are not of interest here. We also used the NASA Task load index.

The 14 items were divided into two sets after session 1 (based on response times). Set 1 were; both items from Thomson and Oppenheimer, one from Finucane and Gullion, the item from Wason and Brooks, the item from Kahneman and Tversky, the item from Wason, and the item from Shafir. The remaining seven items made up set 2.

Approximately three weeks after session 1 participants got either set 1 or 2 again (after having played the Dice task, an information sampling task, not of interest here).

The study was administered in Qualtrics (Qualtrics, Provo, UT).

Analysis

The test – retest score is based on the performance of the 7 items in session 1 that are identical with the items received in session 2.

Results

Participants got on average 49% of the items correct in session 1 and 50% in session 2. Test – retest correlation was Pearson's $\rho = .701$, $p < .001$. Test - retest correlation between all 14 items in session 1 and the seven items in session 2 was $\rho = .8809$, $p < .001$.

Study S2: Test – retest reliability of the Cognitive Effort discounting task

Methods

Ethics

All methods were performed in accordance with the relevant guidelines and regulations and approved by the Institutional Review Board at the Department of Psychology at UiT – The Arctic University of Norway. Written informed consent was obtained from all participants.

Participants

In total 25 participants were recruited, and all completed all three sessions (10 women, 15 men, aged 19-31 years). Inclusion criteria were good health, tolerance for caffeine, sucrose or artificial sweetener as this study served as pilot for a study on the effect of energy drinks on cognitive ability and effort. The participants were compensated for their time with a gift card at their local grocery store valued at 400 NOK (approximately $40).

Materials

The experiment was conducted in English and all instructions were in English. We used the COGED as in experiment 6 with two modifications. In the training phase participants filled out the N-TLX after each n-back level. The instruction slides had a fixed time to ensure that all spent a similar amount of time on the instructions. COGED testing lasted approximately 35 min.

Analysis

We averaged the indifference point (IP) for the choice options 1- vs 2-back, 1- vs-3-back and 1- vs 4-back. Correlation between the averaged IP from session 1 and 2, and session 2 and 3 were calculated.

Results

The average IP was 1.41 in session 1, 1.57 in session 2 and 1.74 in session 3. Participants choose higher n-back levels the more proficient they got with the task. Test – retest correlation was high, Pearson's ρ = .789 for session 1 with session 2, and ρ = .819 for session 2 with session 3.

Additional references

1. Toplak ME, West RF, Stanovich KE. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking & Reasoning. 2014;20(2):147-68. doi: 10.1080/13546783.2013.844729.
2. Levesque HJ. Knowledge Representation and Reasoning. Annual Review of Computer Science. 1986;1(1):255-87. doi: 10.1146/annurev.cs.01.060186.001351.
3. Koehler DJ, James G. Probability matching and strategy availability. Mem Cognit. 2010;38(6):667-76. Epub 2010/09/21. doi: 10.3758/mc.38.6.667. PubMed PMID: 20852231.
4. Teigen KH, Keren G. Waiting for the bus: when base-rates refuse to be neglected. Cognition. 2007;103(3):337-57. Epub 2006/05/26. doi: 10.1016/j.cognition.2006.03.007. PubMed PMID: 16723123.
5. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping Doctors and Patients Make Sense of Health Statistics. Psychol Sci Public Interest. 2007;8(2):53-96. Epub 2007/11/01. doi: 10.1111/j.1539-6053.2008.00033.x. PubMed PMID: 26161749.
6. Toplak ME, Liu E, Macpherson R, Toneatto T, Stanovich KE. The reasoning skills and thinking dispositions of problem gamblers: A dual-process taxonomy. Journal of Behavioral Decision Making. 2007;20(2):103-24. doi: 10.1002/bdm.544.
7. Stanovich KE, West RF. Individual differences in rational thought. Journal of Experimental Psychology: General. 1998;127(2):161-88. doi: 10.1037/0096-3445.127.2.161.
8. Smullyan RM. What is the name of this book? The riddle of Dracula and other logical puzzles. Englewood Cliffs, NJ: Prentice-Hall; 1978.
9. Lehman DR, Lempert RO, Nisbett RE. The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. American Psychologist. 1988;43(6):431-42. doi: 10.1037/0003-066X.43.6.431.
10. Pennycook G, Allan Cheyne J, Barr N, Koehler DJ, Fugelsang JA. On the reception and detection of pseudo-profound bullshit. Judgment and Decision Making. 2015;10(6):549-63. Epub 2023/01/01. doi: 10.1017/S1930297500006999.
11. Treadway MT, Buckholtz JW, Schwartzman AN, Lambert WE, Zald DH. Worth the 'EEfRT'? The effort expenditure for rewards task as an objective measure of motivation and anhedonia. PLoS One. 2009;4(8):e6598. Epub 2009/08/13. doi: 10.1371/journal.pone.0006598. PubMed PMID: 19672310; PubMed Central PMCID: PMCPMC2720457.

12. Xu X, Demos KE, Leahey TM, Hart CN, Trautvetter J, Coward P, et al. Failure to Replicate Depletion of Self-Control. PLOS ONE. 2014;9(10):e109950. doi: 10.1371/journal.pone.0109950.
13. Toplak ME, West RF, Stanovich KE. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory & Cognition. 2011;39(7):1275. doi: 10.3758/s13421-011-0104-1.
14. West RF, Toplak ME, Stanovich KE. Heuristics and Biases as Measures of Critical Thinking: Associations with Cognitive Ability and Thinking Dispositions. Journal of Educational Psychology. 2008;100(4):930-41. doi: 10.1037/a0012842.
15. Van Dooren W, De Bock D, Evers M, Verschaffel L. Students' Overuse of Proportionality on Missing-Value Problems: How Numbers May Change Solutions. Journal for Research in Mathematics Education. 2009;40(2):187-211.
16. Primi C, Morsanyi K, Chiesi F, Donati MA, Hamilton J. The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). Journal of Behavioral Decision Making. 2016;29(5):453-69. doi: https://doi.org/10.1002/bdm.1883.
17. Thomson KS, Oppenheimer DM. Investigating an alternate form of the cognitive reflection test. Judgment and Decision Making. 2016;11(1):99-113.
18. Finucane ML, Gullion CM. Developing a tool for measuring the decision-making competence of older adults. Psychology and aging. 2010;25(2):271-88. doi: 10.1037/a0019106. PubMed PMID: 20545413; PubMed Central PMCID: PMCPMC2918639.
19. Wason PC, Brooks PG. THOG: The anatomy of a problem. Psychological Research. 1979;41(1):79-90. doi: 10.1007/BF00309425.
20. Kahneman D, Tversky A. Subjective probability: A judgment of representativeness. Cognitive Psychology. 1972;3(3):430-54. doi: https://doi.org/10.1016/0010-0285(72)90016-3.
21. Kornreich C, Delle-Vigne D, Brevers D, Tecco J, Campanella S, Noël X, et al. Conditional Reasoning in Schizophrenic Patients. Evolutionary Psychology. 2017;15(3):1474704917721713. doi: 10.1177/1474704917721713. PubMed PMID: 28783973.
22. Shafir E. Uncertainty and the difficulty of thinking through disjunctions. Cognition. 1994;50(1):403-30. doi: https://doi.org/10.1016/0010-0277(94)90038-8.
23. Cokely ET, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R. Measuring risk literacy: the berlin numeracy test. Judgment and Decision making. 2012.
24. Cor MK, Haertel E, Krosnick JA, Malhotra N. Improving ability measurement in surveys by following the principles of IRT: The Wordsum vocabulary test in the General Social Survey. Social Science Research. 2012;41(5):1003-16. doi: https://doi.org/10.1016/j.ssresearch.2012.05.007.

Paper 2

Received: 13 December 2022 | Revised: 13 October 2023 | Accepted: 3 January 2024

DOI: 10.1111/psyp.14532

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Teleological reasoning bias is predicted by pupil dynamics: Evidence for the extensive integration account of bias in reasoning

Martin Jensen Mækelæ[1] | Isabel V. Kreis[1,2] | Gerit Pfuhl[1,3]

[1]Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway

[2]Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[3]Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Martin Jensen Mækelæ, Department of Psychology, UiT The Arctic University of Norway, Tromsø, Norway.
Email: maekelae.m.j@gmail.com

Abstract

Teleological reasoning is the tendency for humans to see purpose and intentionality in natural phenomena when there is none. In this study, we assess three competing theories on how bias in reasoning arises by examining performance on a teleological reasoning task while measuring pupil size and response times. We replicate that humans ($N=45$) are prone to accept false teleological explanations. Further, we show that errors on the teleological reasoning task are associated with slower response times, smaller baseline pupil size, and larger pupil dilations. The results are in line with the single-process extensive integration account and directly oppose predictions from dual-processing accounts. Lastly, by modeling responses with a drift-diffusion model, we find that larger baseline pupil size is associated with lower decision threshold and higher drift rate, whereas larger pupil dilations are associated with higher decision threshold and lower drift rate. The results highlight the role of neural gain and the Locus Coeruleus–Norepinephrine system in modulating evidence integration and bias in reasoning. Thus, teleological reasoning and susceptibility to bias likely arise due to extensive processing rather than through fast and effortless processing.

KEYWORDS

decision-making, drift-diffusion model, dual-process, extensive integration, Locus Coeruleus, neural gain, norepinephrine, pupillometry, reasoning bias, teleological reasoning

1 | INTRODUCTION

Human reasoning and decision-making are prone to bias. A salient example is the tendency to see purpose and intentionality in natural phenomena when there is none. This is known as teleological reasoning (Kelemen et al., 2013). As with other well-documented reasoning biases, what causes this non-normative reasoning remains elusive (Kelemen, 1999). In this paper, we assess three competing theories on how bias in reasoning arises by examining performance on a teleological reasoning task while measuring pupil size and response times.

Teleological reasoning is seen early in children's reasoning development as an explanatory default (DiYanni & Kelemen, 2005). This bias is so persistent that even physical scientists have been shown to endorse false teleological explanations, such as "Trees produce oxygen so that animals can breathe." under time pressure (Kelemen et al., 2013). It is proposed that teleological reasoning remains a cognitive default throughout life (Kelemen et al., 2013). Teleological beliefs may be replaced later in life with scientific normative explanations such as "Oxygen produced by trees is a by-product of photosynthesis." It is not known if this new mindware (scientific explanations) becomes intuitive knowledge for smarter individuals (Raoelison et al., 2020; Stanovich, 2018) or if teleological reasoning always needs to be suppressed by deliberative processing (Evans, 2008; Kahneman, 2011). These two explanations are in line with the Smart intuitor and Default-Interventionist dual-process models, respectively, which have been highly influential in research on bias in reasoning (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011; Pennycook et al., 2015; Stanovich, 2009a, 2009b). We here briefly introduce two dual-process models, the Default-Interventionist account and the Smart intuitor account. Alternative dual-process models were not included as they failed to make clear and distinct predictions from the Default-Interventionist and Smart intuitor accounts in this task (Epstein, 1994; Sloman, 1996).

1.1 | Dual-process models

At the core, dual-processing accounts state that human reasoning can be separated into two different modes of processing (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011). Type 1 processing, often called intuitive or heuristic, is automatic and does not require working memory capacity, that is, measurable features of Type 1 processing are being fast and effortless. Type 2 processing, often called analytic or deliberate, relies on working memory resources and uses mental simulation to generate responses. Measurable features of Type 2 processing are being slow and effortful. Accordingly, these processes can be gauged by measuring response times and pupil dilations, as the pupil is known to dilate with increasing cognitive effort (Hess & Polt, 1964; Kahneman & Beatty, 1966; van der Wel & van Steenbergen, 2018).

1.2 | Default-interventionist account

The Default-Interventionist account (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011) proposes that Type 1 processes are the default. Type 2 processes are engaged at later stages of reasoning, or not at all. The Default-Interventionist account proposes that humans are cognitive misers because their default is to conserve effort expenditure by relying on Type 1 processing. Thus, bias in reasoning occurs due to overreliance on fast effortless Type 1 processing and failure to engage in slow, effortful Type 2 processing when called for. According to the Default-Interventionist account, an intuitive teleological explanatory default produced by Type 1 processes (e.g., "Trees produce oxygen so that animals can breathe.") would have to be inhibited and overridden by Type 2 processing to produce a normative scientifically accurate explanation (e.g., "Oxygen is a by-product of photosynthesis./Trees do not produce oxygen so that animals can breathe.") when trying to understand events and phenomena. Importantly, the Default-Interventionist account predicts that overriding a false teleological explanation would require longer response times and more effort, compared to accepting a false teleological explanation which should be fast and effortless.

1.3 | The smart intuitor account

The Smart intuitor account has evolved from the Default-Interventionist account as an increasing number of studies show evidence opposing predictions from the Default-Interventionist account (Bago & De Neys, 2017, 2019; Newman et al., 2017; Raoelison et al., 2020; Raoelison & De Neys, 2019; Thompson et al., 2011). An example of this was shown by Raoelison et al. (2020) with a two-response paradigm for the cognitive reflection test (Frederick, 2005). The cognitive reflection test has been developed to assess an individual's ability to override an initial intuitive incorrect response in order to produce a deliberate correct response (consistent with Default-Interventionist account). However, Raoelison et al. (2020) showed that most correct responses were made fast (intuitively), and very few correct responses were due to respondents' initial wrong response followed by a correction after deliberation. Accordingly, the Smart intuitor account proposes that Type 1 processing can produce many types of intuitions which were previously believed could only arise from Type 2 processing (Bago & De Neys, 2019; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996; Thompson et al., 2018). Importantly, the Smart intuitor account proposes that high cognitive capacity individuals are more likely to answer correctly on reasoning tasks by having "better" or more accurate intuitions (Bago & De Neys, 2017, 2019; Raoelison et al., 2020). A corrective deliberate process (as proposed by Default-Interventionist) can still happen, but most correct responses in decision-making tasks are due to accurate intuitions rather than overriding faulty intuitions (Raoelison et al., 2020). The Smart intuitor account predicts then that overriding of false teleological explanations is not always necessary. Both teleological intuitions and scientifically normative intuitions can be made intuitively through a fast and effortless Type 1 process. More generally, the Smart intuitor account predicts that both normative responses and errors can be made fast and with little effort. However, when engaging in Type 2 processing, seen by longer response times and more effort, the normative response is more likely.

To distinguish between the Default-Interventionist and Smart intuitor accounts, we included individual difference measures of cognitive ability and cognitive motivation. According to the Default-Interventionist account, engaging in Type 2 processing increases the probability of normative responses. Therefore, performance on the teleological reasoning task should be associated with higher trait cognitive motivation (Cacioppo et al., 1996; Stanovich, 2009b; Toplak et al., 2011, 2014; West et al., 2008). However, if normative responses are made intuitively by individuals high in cognitive ability as proposed by the Smart intuitor account, then cognitive ability should be associated with performance and cognitive motivation should have less influence on normative responding (Raoelison et al., 2020).

Importantly, underlying both the Default-Interventionist and Smart intuitor account is the assumption that more effortful and extensive processing (Type 2) leads to more normative responses and less bias. However, a single-process framework, the Extensive integration account, makes the opposite prediction, namely that bias in reasoning is exacerbated by more extensive processing. Recently, Eldar et al. (2021) highlighted that dual-process theories and the Extensive integration account make opposing predictions regarding pupil dilation and found support for the Extensive integration account in three framing tasks.

1.4 | Extensive integration, neural gain, and the locus coeruleus–norepinephrine system

The Extensive integration account builds on a single-process framework where decision-making is seen as a dynamic process of gradual noisy evidence accumulation and integration leading up to a decision (Busemeyer et al., 2006; Busemeyer & Townsend, 1993; Krajbich & Rangel, 2011; Usher et al., 2013; Usher & McClelland, 2004). Here, bias accumulates if the decision-making process unfolds over many time steps. Thus, a small bias will have larger effects if each piece of evidence has lower weighting and the decision requires a longer evidence accumulation process. Thus, more extensive integration is associated with more bias (Eldar et al., 2021; Usher & McClelland, 2004). Importantly, it is proposed that evidence integration is influenced by the Locus Coeruleus–Norepinephrine system, as norepinephrine modulates neural gain (Aston-Jones & Cohen, 2005; Eldar et al., 2013; Eldar, Cohen, et al., 2016; Eldar, Niv, et al., 2016; Jepma & Nieuwenhuis, 2011; Joshi et al., 2016). Low neural gain leads to lower weighting of each piece of evidence, and thus more extensive integration is required to reach a decision (Eldar, Cohen, et al., 2016; Eldar et al., 2013, 2021). Conversely, high neural gain leads to increased weighting of each piece of evidence. Importantly, neural gain can be gauged with pupillometry as pupil diameter is highly correlated with Locus Coeruleus activity (Aston-Jones & Cohen, 2005; Eldar et al., 2021; Gilzenrat et al., 2010; Reimer et al., 2016). Smaller baseline pupil diameter indicates low tonic Locus Coeruleus activity, low norepinephrine levels, and low neural gain (Aston-Jones & Cohen, 2005; Berridge & Waterhouse, 2003; Eldar et al., 2013; Eldar, Niv, et al., 2016; Joshi et al., 2016; Reimer et al., 2016). Additionally, larger pupil dilations can also indicate low neural gain as baseline pupil size and baseline-corrected pupil dilations are inversely correlated (Aston-Jones & Cohen, 2005; Eldar et al., 2013; Gilzenrat et al., 2010). Thus, according to the Extensive integration account, bias in reasoning occurs due to more extensive evidence integration, which is exacerbated by low neural gain. Therefore, the Extensive integration account predicts that biased responses (i.e., teleological reasoning errors) are associated with longer response times and larger pupil dilations (indicating low neural gain).

In this study, we assessed which of the three accounts best explains teleological reasoning bias by evaluating performance on a teleological reasoning task. A teleological reasoning bias is evident if participants make more errors when evaluating the truth of false teleological explanations compared to comparable control statements (such as physical explanations and true teleological explanations, see methods). Both dual-process models predict that slower response times and larger pupil dilations are associated with more normative responses, that is, rejecting false teleological explanations (e.g., "Trees produce oxygen so that animals can breathe"). The Extensive integration account makes opposing predictions, namely that normative responses are associated with fast responses and smaller pupil dilations. Additionally, the Extensive integration account predicts that larger baseline pupil size is associated with normative responding.

Table 1 summarizes the predictions across the three accounts.

1.5 | Exploratory analyses and pre-registration

As an exploratory measure we recorded pupil dilations following feedback (correct or incorrect) that

TABLE 1 Predictions of the three accounts for responses in the teleological reasoning task.

Parameter	Default-interventionist	Smart intuitor	Extensive integration
Response time	Slow responses are more likely normative. Fast responses are more likely errors	Slow responses are more likely normative. Fast responses can be both normative and errors	Fast responses are more likely normative. Slow responses are more likely errors
Pupil dilation	Larger dilations are more likely normative responses. Smaller dilations are more likely errors	Larger dilations are more likely normative responses. Smaller dilations can be both errors and normative responses	Smaller dilations are more likely normative responses. Larger dilations are more likely errors
Baseline pupil size	N/A	N/A	Larger baseline more likely leads to normative responses
Cognitive ability	High ability predicts better performance (but see Stanovich and West [2008])	High ability predicts better performance	N/A
Cognitive motivation	High cognitive motivation predicts better performance	Cognitive motivation has less impact on performance than cognitive ability	N/A

Note: Predictions where the three accounts make similar predictions are not included, for example, pupil dilation to feedback (see text).

participants received after their responses in the teleological reasoning task. Pupil dilation has been linked to decision uncertainty and the following surprise after feedback (Colizoli et al., 2018; de Gee et al., 2021; Preuschoff et al., 2011; Urai et al., 2017). We expected larger pupil dilation, signaling surprise, for error trials compared to trials with correct responses. Further, we expected larger pupil dilation where decision confidence was high, but the feedback indicated being incorrect, and smaller dilations on trials where decision confidence was low. Pupil dilation to feedback cannot confirm or disconfirm any account.

Lastly, in accordance with the Extensive integration account, we modeled responses on the teleological reasoning task with an established sequential sampling model of the decision process, the drift-diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008; Smith & Ratcliff, 2004). The drift-diffusion model allows for the investigation of latent psychological processes underlying decisions (Ratcliff & McKoon, 2008; Wiecki et al., 2013). Additionally, the drift-diffusion model enables investigation of the link between psychological processes and neural mechanisms by utilizing physiological measures (i.e., pupil dilation) as predictors of parameters in the drift-diffusion model (Cavanagh et al., 2011, 2014; Wiecki et al., 2013).

Pre-registration for this study is available on OSF (https://osf.io/vk7r4/). Our pre-registered hypotheses were in line with the Default-Interventionist dual-process account. Please note, we deviate from the pre-registration as the analysis plan was found to be inadequate. Additionally, the pre-registration included plans to assess heart-rate variability; however, due to low-quality recordings (Empatica E4), these data could not be analyzed and are hence not described further.

2 | METHODS

2.1 | Participants

Participants were non-psychology students, $N=45$ (27 female), and mean age was 23.35 years (range 18–37). Participants reported not having any neurological disorder, history of brain disease or surgery, and not taking any central nervous system medication or drugs. In addition, as all test stimuli were in English and participants had different native languages, self-rated English proficiency had to be higher than 4 on a scale from 1 to 7, where 1 = "understand a few words" and 7 = "Master it like native language". The threshold was set based on a previous study showing no difference in deliberate reasoning performance between native and second language, and no effect of English proficiency on deliberate reasoning for participants scoring above 4 on the same English proficiency scale (Mækelæ & Pfuhl, 2019). All participants gave written informed consent prior to participation. The study was approved by the institutional review board at the Department of Psychology, UiT, The Arctic University of Norway. Participants received a voucher worth 400 NOK (approximately 40 USD) for participating in two test sessions (from test session two we included two cognitive ability measures in the SOM where we report the relationship between performance on the teleological reasoning task and two cognitive ability measures).

2.2 | Materials

2.2.1 | Cognitive motivation

We used the 18-item Need for Cognition Scale (NFC) (Cacioppo et al., 1984), which measures a person's tendency to engage in and enjoy cognitively effortful activities. An example item is "*I prefer complex to simple problems*." The 18 items are rated on a 5-point Likert scale from 1 = "*Extremely uncharacteristic of me*" to 5 = "*Extremely characteristic of me*." Total score can range from 18 to 90. Internal consistency was high, McDonalds $\omega = 0.86$. The scale was implemented in Qualtrics (Qualtrics, Provo, UT).

2.2.2 | Cognitive ability

We used a composite of rational reasoning tasks to measure cognitive ability. The battery of rational reasoning tasks was created with 14 items from the heuristics and biases literature. We used items 2–7 from the Cognitive Reflection Test (Toplak et al., 2014), one fully disjunctive reasoning problem; "the marriage problem" (Levesque, 1986), one probability matching task (Koehler & James, 2010), one probability estimation task; "the bus problem" (Teigen & Keren, 2007), one making sense of medical results problem (Gigerenzer et al., 2007), one Bayesian reasoning problem (Toplak et al., 2007), adapted from Fischhoff and Beyth-Marom (1983), one covariation detection problem (Stanovich & West, 1998), one knight and knave problem (Smullyan, 1978), and one conditional reasoning problem (Lehman et al., 1988). Correct answers were scored as 1, incorrect as 0. Total composite rational reasoning score ranged between 0 and 14. The task was implemented in Qualtrics (Qualtrics, Provo, UT).

2.2.3 | Teleological reasoning

The teleological reasoning task consisted of statements containing false teleological explanations (test items), as well as control statements (control items) that participants were asked to judge as true or false (Kelemen et al., 2013; Kelemen & Rosset, 2009). There were 77 items in total, 34 of which were test items consisting of false teleological explanations for natural phenomena (e.g., "Trees produce oxygen so that animals can breathe."). The 43 control items consisted of 24 physical explanations that were either true ("Objects fall downwards because they are affected by gravity.") or false ("Soup is hot because it is primarily liquid."), and 19 control teleological explanations that were either true ("Schools exist in order to help people learn new things.") or false ("Mice run away from cats in order to get exercise."). Thus, test sentences are false teleological explanations in the domain of natural phenomena where the stated explanations are inappropriate. Control sentences are teleological explanations concerning the social–conventional and artifact domains where these explanations are appropriate.

The task was computerized with stimulus sentences presented auditorily via noise-canceling headphones. The task was self-paced, and each trial was initiated by pressing the space bar. Trials started with a fixation cross appearing on screen, and the auditory stimulus was presented after a delay of 0.5 s (see Figure 1). Stimulus sentences varied in duration between 2.3 and 3.7 s. After the stimulus sentence ended, participants had 4 s to respond, indicating whether the statement was true or false by pressing "D" or "K" on a QWERTY keyboard, respectively. Participants received feedback 1.8–2.4 s after their answer, by a "V" or "X" appearing in place of the fixation cross (feedback duration 4.0–6.2 s, uniformly jittered), representing correct and incorrect responses, respectively. If a participant did not respond within the 4 s, the trial was amended to the end of the task for repetition. All stimuli presented on screen were isoluminant. Items were pseudo-randomized with the constraint of not more than three in a row of the same type (test items or control items).

Instructions, fixation cross, and feedback for the task were presented on a monitor (width 34 cm, height 27 cm, resolution 1280 × 1024). The teleological reasoning task was programmed in Python (version 3.7) and presented in Psychopy (Peirce et al., 2019), script available on OSF (https://osf.io/vk7r4/). The auditory test stimuli for the teleological reasoning task were created by entering the

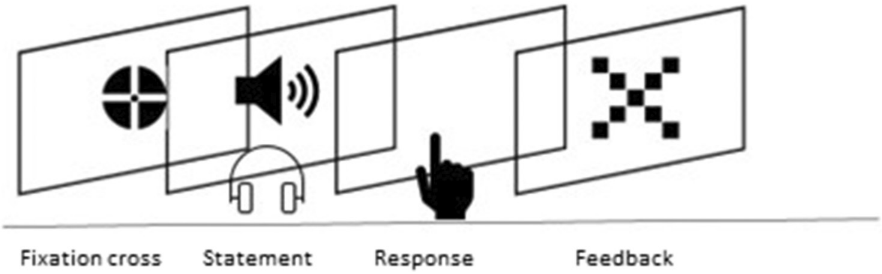


FIGURE 1 Teleological reasoning task. Trial structure of the Teleological reasoning task. Fixation cross, duration 200 ms. Statement (stimulus onset delayed by 0.5 s) presented auditorily (length 2.3–3.7 s). Feedback (onset 1.8–2.4 s after response, jittered) indicating correct and incorrect responses (here X for being wrong) presented on screen (4.0–6.2 s). Figures not to scale.

stimulus statements into Google Cloud 's speech-to-text API (Demo provided by Google Cloud online (available at: https://cloud.google.com/speech-to-text, [accessed 08.27.2019])). The resulting output was recorded with the audio recording and editing software Audacity® (Version 2.3.2, Audacity Team, 2019), audio files are available on OSF (https://osf.io/vk7r4/).

2.3 | Pupil recording

Pupil size was recorded during the teleological reasoning task with a desk-mounted Eyelink 1000 eye tracker (SR-Research, Ontario, Canada) with a sampling rate of 500 Hz. A chinrest was used to stabilize head position and viewing distance (65 cm from top of screen, 69 cm from screen bottom). A two-minute baseline measurement of pupil dilation was recorded in a sitting position in front of the computer before the teleological reasoning task started. Participants were instructed to fixate on the center of the screen.

2.3.1 | Procedure

Participants were recruited through flyers at UiT, The Arctic University of Norway. Participants were individually tested by a trained experimenter. The order of the tasks were cognitive ability, cognitive motivation, and teleological reasoning task. The test session included assessments for a separate replication project (Mækelæ et al., 2023), with a Demand selection task (Kool et al., 2010) followed by NASA task load index (Hart & Staveland, 1988) and another N-TLX assessment following cognitive ability, administered at the beginning of the session in a different room. However, these assessments are not relevant to the current study and were not expected to affect performance in any of the other assessments.

2.4 | Data processing

Data processing of pupil measurement was performed in the statistical environment R (version 4.1.2.) (R Core Team, 2021). Eyeblinks and other artifacts (rapid changes in pupil size, caused by head movements, lid flickering, etc.) were detected based on the signal's velocity (Mathot, 2018) and corrected using linear interpolation. Here, thresholds and on- and offset margins for the interpolation window were adapted on an individual basis, due to inter-individual differences in signal recovery (the speed at which the signal returns back to normal). The interpolated signal was smoothed with a 3 Hz low-pass Butterworth filter. If blinks or artifacts spanned more than 1000 consecutive milliseconds, the respective interpolated signal was treated as missing. Finally, the signal was visually screened, and trials with remaining artifacts were identified and excluded from further analysis if the artifacts occurred during time windows of interest (trial baseline, decision, and feedback; $n = 0.5$ trials per participant on average). For each trial, baseline pupil size ("Baseline pupil") was calculated as the average signal across the first 200 ms following the onset of the fixation cross. Pupil dilation during decision-making, that is, the time window from onset of the auditory stimulus until response made, and during feedback processing, that is, the time window between feedback onset and the subsequent 3000 ms, was baseline-corrected by subtracting baseline pupil from every sample within the respective time window of interest. Maximum pupil dilation during decision-making ("PDmax-BL") and during feedback ("Feedback PDmax-BL") were extracted. For decision-making, maximum pupil dilation was further calculated based on the raw signal, without prior baseline correction ("PDmax").

Baseline pupil, PDmax, PDmax-BL, and Feedback PDmax-BL measures were treated as missing (NA) if more than 50% of the signal within the respective time window were missing and/or interpolated.

2.5 | Data analyses

Linear mixed models were analyzed with the lme4 package (Bates et al., 2015). Modeling of responses on the base-rate tasks with the drift-diffusion model was performed with Python (version 3.9) (Patil et al., 2010). The model was implemented with the hierarchical drift-diffusion model, contained in the dockerHDDM (Pan et al., 2022; Wiecki et al., 2013).

First, we aimed to replicate that humans show a teleological reasoning bias. We assessed whether false teleological explanations (test condition) lead to more errors in reasoning compared to comparable explanations (control condition) by testing if there was a significant difference in accuracy between the test and control conditions.

Second, to investigate which of the three accounts best explains performance in the teleological reasoning task we applied separate generalized linear mixed models (GLMM) for response times and pupil dilations. All reported models successfully converged. We only report relevant estimates of fixed factors in the manuscript; for more details on the models, see SOM (Tables S1–S5 and S7–S10). For the pupil analysis, the main analysis is conducted with maximum pupil dilation with Baseline pupil subtracted ("PDmax-BL") as this is a common way to report pupil dilation (Mathot, 2018), also referred to as phasic response. Further, we report analyses with Baseline pupil (also referred to as tonic response) and PDmax (uncorrected) entered separately as this is of particular interest for the Extensive integration account. We note that the latter approach may lead to multicollinearity issues; however, centering of the variables alleviates this. Assessment of variance inflation factor with the "caret" package (Kuhn, 2015) and visual inspection of the residuals with the "DHARMa" package (Hartig, 2022) showed no multicollinearity issues.

Third, to investigate how individual differences in cognitive ability and cognitive motivation influence susceptibility to false teleological explanations, we performed a linear model with cognitive motivation and cognitive ability as predictors of accuracy in the test condition.

Values for "Baseline pupil", "PDmax", "PDmax-BL", and "Feedback PDmax-BL" were separately *z*-scored within participants. Cognitive motivation and cognitive ability were *z*-scored across participants.

2.6 | Exploratory analyses

Pupil dilations following feedback were recorded to investigate uncertainty and surprise in the teleological reasoning task. We applied a Linear mixed model (LMM) with Feedback PDmax-BL as outcome with condition and accuracy as fixed factors.

A drift-diffusion model was applied to investigate latent psychological processes underlying decision in the teleological reasoning task and the influence of pupil dynamics. The drift-diffusion model is an established computational model of the decision process consistent with the Extensive integration account (Ratcliff, 1978; Ratcliff & McKoon, 2008; Smith & Ratcliff, 2004). We note that the drift-diffusion model was accuracy coded, meaning the decision boundaries are correct and incorrect responses, and accordingly do not include a bias parameter.

First, we assessed whether there was a difference in the decision process when evaluating false teleological explanations compared to control statements, by testing if there were significant differences in the main parameters of the drift-diffusion model in the test and control condition. Second, pupil data were applied as a linear predictor of trial-by-trial variation in drift rate, threshold, and drift-rate variability. We ran the analyses with both "PDmax-BL" and separately entered "PDmax" and "Baseline pupil" as predictors.

For each model, we ran five Markov chains with 20,000 samples each, 12,000 of which were burn-in. Every second sample was discarded as thinning in order to reduce autocorrelation in chains. Model convergence was assessed with visual inspection of the trace, autocorrelation, the marginal posterior, and the Gelman-Rubin R statistic. All parameters had an R-hat value below 1.01. Model comparison was conducted with the deviance information criterion (DIC). Lower DIC indicates better fit. However, we note results of models with fit in similar range as DIC has limitations when comparing fit. See SOM Table S11 for comparison of all models.

2.7 | Sample size

Our sample size rationale was based on a comparable study linking pupil responses to prediction-making in environments with changing stochastic structure (de Berker et al., 2016; Kreis et al., 2023). In this study, a pupillary sensitivity measure to uncertainty correlated highly positively with performance (Pearson correlation coefficient $r = .62$, $n = 22$). Assuming some regression to the mean, we based our sample size calculation on a smaller effect size, $r = .4$, α of 0.05 (two-sided test), power of 0.8, which yielded 44 participants in the analysis (G power 3.1). Regarding individual differences, Thompson et al. (2018) report large effect sizes (η^2 of 0.3 to 0.6), and thus a sample of 40 participants would be sufficient to find an effect. Our final sample after exclusions was deemed sufficient to continue with analyses.

3 | RESULTS

A total of six participants were excluded, two by their behavioral responses (one failed to respond, one mixed up buttons), and four had too low quality of their pupil data or calibration failed, leaving a total of 39 participants (see SOM for behavioral analysis prior to exclusions by low-quality pupil data, i.e., with $n = 43$).

Descriptive statistics for all variables can be found in Table 2.

3.1 | Accuracy

To assess if participants showed a teleological reasoning bias, we compared participants' performance in the test condition to the control condition. A Mann–Whitney U test showed that the percentage of correct responses in the control condition (Mdn = 91.9, *SD* = 5.4) was significantly higher than the percentage of correct responses in the test condition (Mdn = 75.0, *SD* = 14.4), U = 1315, $p < .001$. This indicates that participants on average showed a teleological reasoning bias and endorsed false teleological explanations.

TABLE 2 Descriptive statistics.

	Mean	*SD*	Minimum	Maximum
Baseline pupil (tonic response)	32.81	4.99	21.92	51.59
PDmax	35.73	5.65	24.37	58.49
PDmax-BL (phasic response)	2.91	1.93	−1.54	14.87
Feedback PDmax-BL	1.83	2.33	−11.92	14.29
Response time in seconds	1.21	0.80	0.01	3.96
Cognitive ability	7.21	2.48	3.00	13.00
Cognitive motivation	55.10	10.19	24.00	74.00

Note: Variables not *z*-scored.

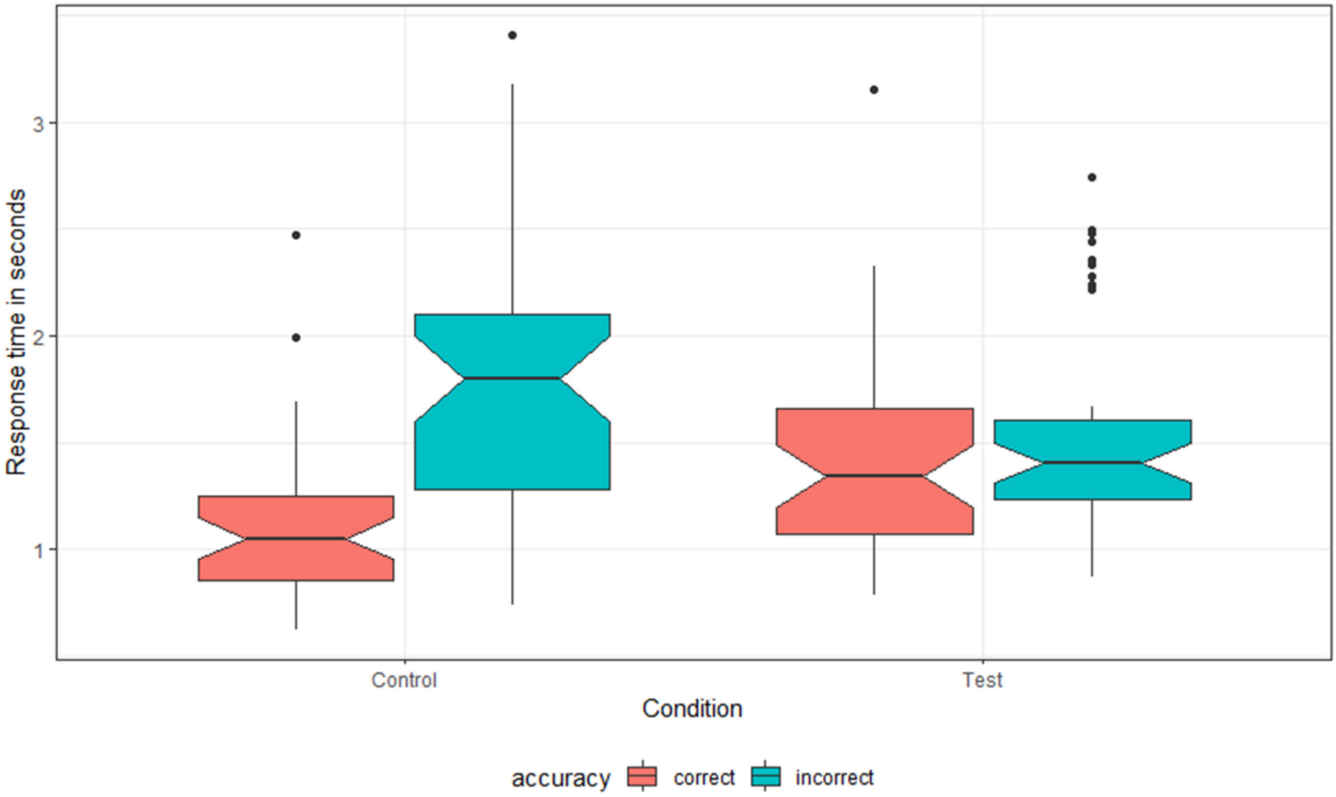


FIGURE 2 Response times separated by condition and accuracy. Response times, average per participant in seconds for the teleological reasoning task. Responses are separated by condition and accuracy.

3.2 | Response times

To assess if normative responses were associated with longer (as predicted by dual-process models) or shorter (as predicted by the Extensive integration account) response times, we applied a GLMM with accuracy as outcome (normative–error responses) and *z*-scored response times and condition as fixed factors and participants as random factors.

The results showed that correct responses were associated with shorter response times ($\beta = -0.48$, $SE = 0.06$, $z = -8.59$, $p < .001$), and that more errors were made in the test condition ($\beta = -1.18$, $SE = 0.13$, $z = -9.28$, $p < .001$), see Figure 2.

3.3 | Pupil dilation – Decision

The most important question in this study is whether errors in teleological reasoning are associated with small or large pupil dilations. The Default-Interventionist account predicts that errors occur through a fast effortless process and would therefore be associated with smaller pupil dilations. The Smart intuitor account predicts that both errors and normative responses can be associated with small pupil dilations; however, if pupil dilations are large, the account predicts that normative responses are more likely. The Extensive integration account, on the other hand, predicts that errors should be associated with larger pupil dilations.

To test if larger or smaller pupil dilations were predictive of correct responses on the teleological reasoning task, we applied a GLMM with accuracy as outcome, PDmax-BL (phasic response) and condition as fixed factors and by-participant random intercepts (see SOM Table S6 for analysis with pupil dilation and effort as outcome).

The results showed that smaller pupil dilations were a significant predictor of normative responses ($\beta = -0.19$, $SE = 0.06$, $z = 3.15$, $p = .002$), and that participants made more errors in the test condition ($\beta = -1.30$, $SE = 0.12$, $z = -10.51$, $p < .001$). Thus, the results indicate that errors are associated with larger pupil dilations (i.e., larger phasic responses). Figure 3 shows average pupil waveform for correct and incorrect responses (see also, SOM Figure S1 for phasic response (z-scored PDmax-BL) in the time window from stimulus sentence onset until response).

Next, the Extensive integration account specifically predicts that lower baseline pupil size and larger pupil dilations are associated with more bias and thus more incorrect responses. To assess the contribution of both Baseline pupil and PDmax, we applied a GLMM with accuracy as outcome, Baseline pupil, PDmax, and condition as fixed factors and by-participant random intercepts.

The results showed that higher Baseline pupil was associated with more correct responses ($\beta = 0.24$, $SE = 0.08$, $z = 3.06$, $p = .002$). Conversely, larger PDmax were associated with more errors ($\beta = -0.21$, $SE = 0.08$, $z = -2.76$, $p = .006$), and the test condition was associated with more errors ($\beta = -1.31$, $SE = 0.12$, $z = 10.53$, $p < .001$). The results showed that errors in teleological reasoning are associated with smaller baseline pupil size (tonic response) and larger pupil dilations (phasic response).

3.4 | Individual differences

To distinguish between the Default-Interventionist and Smart intuitor account, we included individual difference measures of cognitive ability and cognitive motivation. According to the Default-Interventionist account, engaging in Type 2 thinking, thus increasing probability of normative responses, is related to trait differences in cognitive motivation. However, if normative responses are made intuitively by individuals high in cognitive ability as proposed by the Smart intuitor account, then cognitive motivation should make little difference in normative responding.

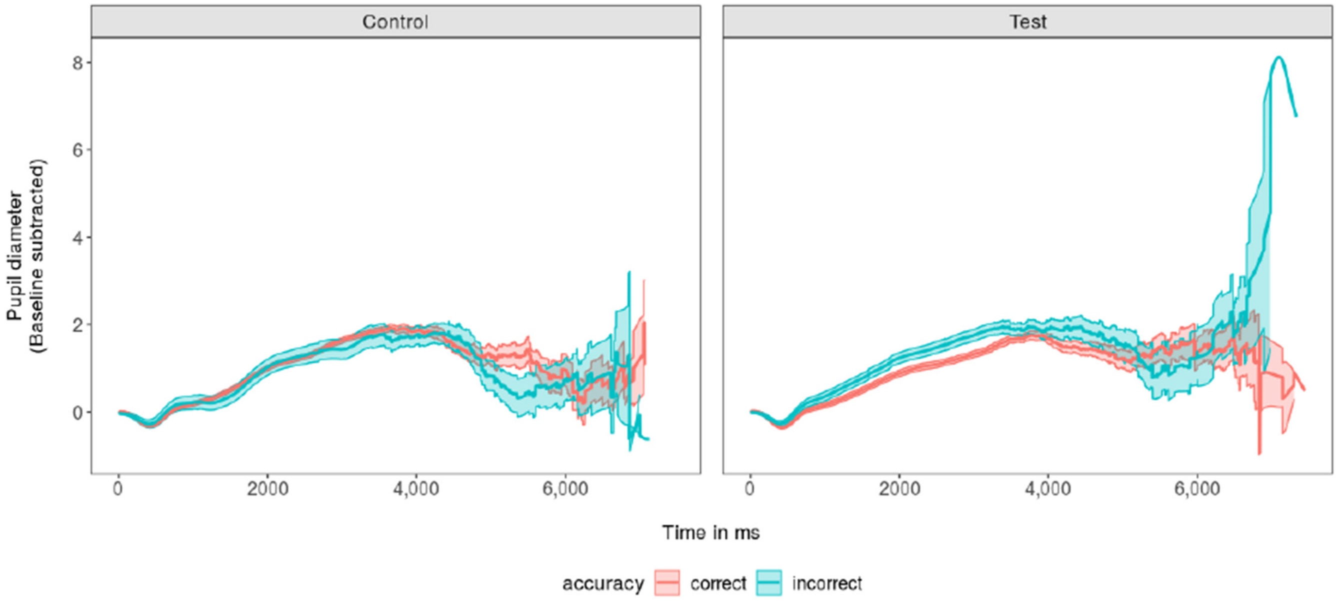


FIGURE 3 Pupil waveform for correct and incorrect responses in the control and test conditions during listening and until a response was made. Change in pupil waveform from onset of the statement until a response was made in the teleological reasoning task. Minimum duration is 2.4 s (shortest statement and immediate responding), maximum is 7.7 s. Pupil waveform is averaged across all participants and trials. Exclusions applied. Shaded area represents standard error. ms, milliseconds.

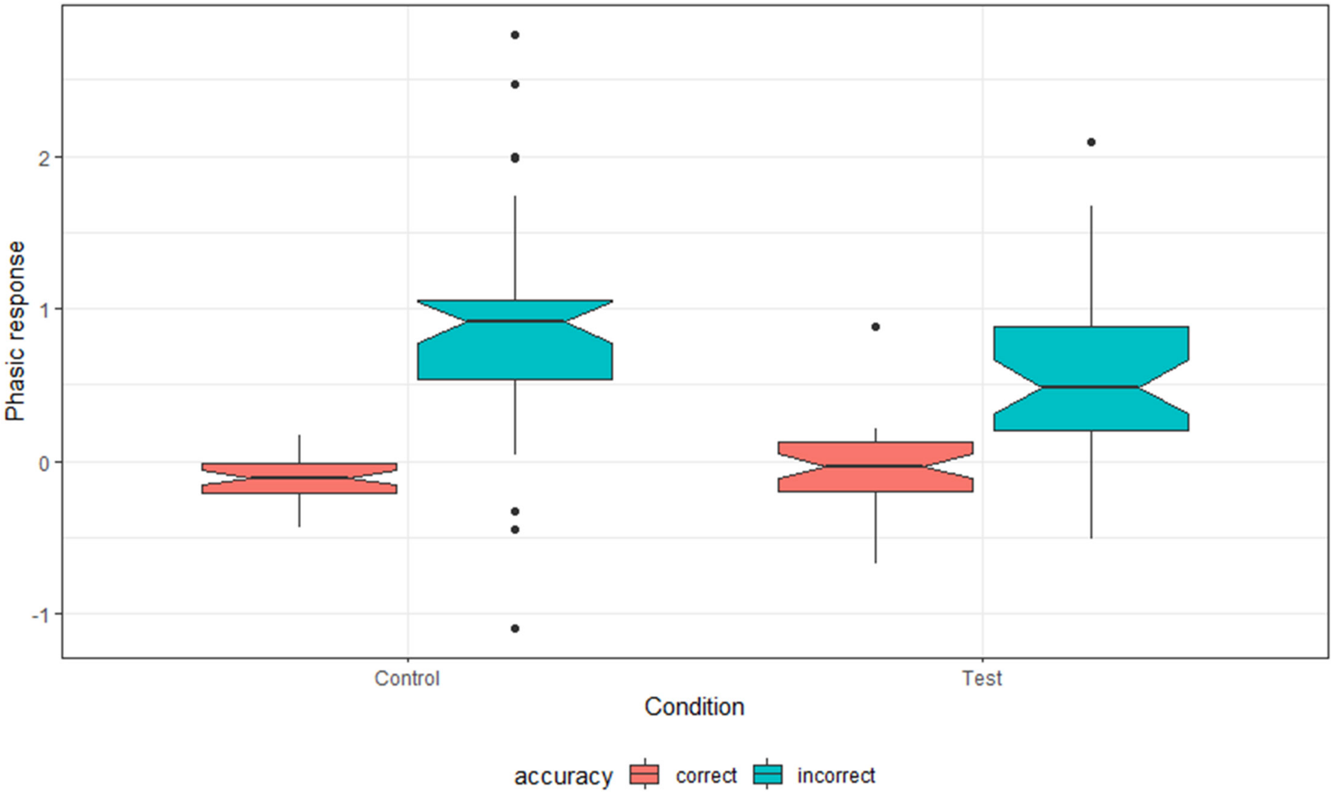


FIGURE 4 Phasic response (z-scored maximum pupil dilation with baseline subtracted) during feedback. Phasic response during feedback is the z-scored maximum pupil dilation with baseline subtracted and averaged per participant. Responses are separated by condition and accuracy.

To investigate how individual differences in cognitive motivation and cognitive ability influence performance, we conducted a linear model with cognitive motivation and cognitive ability as predictors of accuracy in the test condition. The model explained 28.1% of the variance in accuracy, with cognitive ability ($\beta = 0.08$, $SE = 0.02$, $t = 3.74$, $p = .001$) but not cognitive motivation ($\beta = -0.01$, $SE = 0.02$, $t = -0.38$, $p = .710$) as a significant predictor of performance in the test condition. The results show that higher cognitive ability, but not higher cognitive motivation, is associated with successfully rejecting false teleological explanations.[1]

3.5 | Exploratory analyses

3.5.1 | Pupil dilation to feedback

As an exploratory investigation we looked at pupil dilation following feedback, as pupil dilation has been known to signal decision uncertainty and surprise after feedback (de Gee et al., 2021). We interpret large pupil dilations here to indicate more surprise (see Figure 4).

To assess decision uncertainty and surprise for errors and normative responses in the two conditions, we conducted a linear mixed model with Feedback PDmax-BL as outcome and response, and condition and their two-way interaction as fixed factors and by-item[2] random intercepts.

The results yielded a significant interaction ($\beta = 0.36$, $SE = 0.12$, $t = 2.89$, $p = .004$), that is, pupil dilation was largest for incorrect responses in the control condition and smallest for correct responses in the control condition. On average, correct responses were associated with smaller pupil dilations to feedback ($\beta = -0.84$, $SE = 0.10$, $t = -8.58$, $p < .001$) compared to incorrect responses, and pupil dilations were on average larger in the control condition ($\beta = -0.31$, $SE = 0.11$, $t = -2.73$, $p = .006$) compared to the test condition. The result from the analyses of pupil dilation to feedback showed larger pupil dilations for errors, and this effect was larger in the control condition than in the test condition.

[1] SOM contains analysis for two additional measures of cognitive ability for a sub-sample of participants which participated on a separate day for a separate project.

[2] By-item random intercepts were applied as the model failed to converge when including by-participant random intercepts.

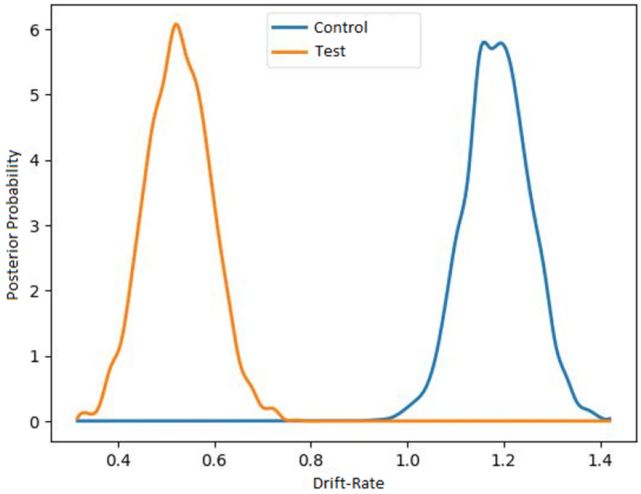


FIGURE 5 Posterior estimate of group mean drift rate in the test and control condition. Significant difference in posterior estimates of group mean drift rate in the test and control condition in the Teleological reasoning task.

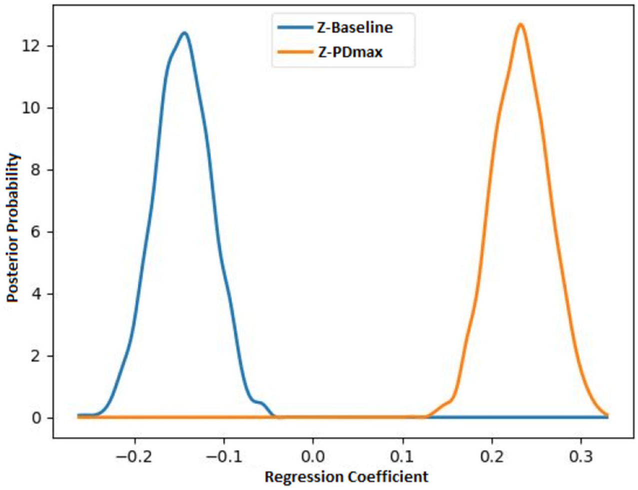


FIGURE 6 Effect of *z*-scored baseline pupil and PDmax on decision threshold. Posterior estimates of regression coefficients for *z*-scored trial-baseline pupil size and *z*-scored maximum pupil dilation as predictors of trial-by-trial variation in threshold.

3.5.2 | Drift-diffusion model

To find the model with the best fit, we analyzed the models in two steps. First, we assessed whether any of the main parameters of the drift-diffusion model differed between the test and control condition. In the second step, we assessed whether pupil measures could predict trial-by-trial variation in parameters of the drift-diffusion model.

In the first step, we found that drift rate was significantly lower in the test condition compared to the control condition (probability of drift rate in test condition being larger than mean in control = 0.01). Posterior estimates of drift rate in test and control conditions can be seen in Figure 5. Threshold was not significantly different in the two conditions (although, near significance level for the threshold being higher in the test condition), with a 0.077 probability of the mean threshold in the test condition being higher than the mean threshold in the control condition (see SOM Figure S2).[3]

In the second step, we applied pupil measures as predictors of trial-by-trial variation in parameters of the drift-diffusion model. According to the Extensive integration account, lower baseline pupil size, and thus also larger pupil dilations (as they are inversely correlated), should be linked to more extensive integration. More extensive integration in the drift-diffusion model can be achieved from either decreased drift rate (lower rate of accumulation toward decision boundary) or increased threshold (response caution) or both.

The winning model indicated by lowest DIC value was the model with *z*-scored Baseline pupil and *z*-scored PDmax as predictors of threshold, with separate drift rate by condition. As can be seen from Figure 6, Baseline pupil and PDmax had opposite effects on the decision threshold. Higher Baseline pupil was linked to lower threshold, whereas higher PDmax was associated with higher decision threshold.

We note that the winning model (Figure 6, DIC = 6088) showed only slightly better fit compared to the model with PDmax-BL as a predictor of drift rate (DIC = 6098) and the model with Baseline pupil and PDmax as predictors of drift rate (DIC = 6100). Importantly, the effect of pupil measures on drift rate was opposite to the effect these measures had on threshold (see SOM Figure S3). That is, higher PDmax-BL was associated with both lower drift rate and higher decision threshold (see SOM Figures S4 and S5). Posterior predictive modeling supported that PDmax-BL as a predictor of threshold had slightly better fit compared to PDmax-BL as a predictor of drift rate (see SOM Figures S7 and S8). Lastly, PDmax-BL was not related to drift-rate variability (see SOM Figure S6).

4 | DISCUSSION

The purpose of this study was to investigate theoretical frameworks that explain bias in reasoning, in particular, teleological reasoning. The participants in the study did show a teleological reasoning bias, as evidenced by their acceptance of false teleological explanations for natural phenomena at a significantly higher rate compared to errors made on comparable control statements. This is

[3] Including drift-rate variability to the model or both separate threshold and drift rate was evaluated as not adding significant improvement to the model.

consistent with previous studies on teleological reasoning (Kelemen et al., 2013). By modeling responses with a drift-diffusion model, we found further support for false teleological explanations being harder to evaluate as the test condition yielded a lower drift rate.

Errors in reasoning were associated with slower response times and larger pupil dilations. Further, smaller baseline pupil size and larger pupil dilations were associated with errors in reasoning. Thus, the results strongly support the extensive integration account of bias in reasoning and provide no support for dual-processing accounts.

The extensive integration account relies on a framework where decision-making is seen as a noisy sequential sampling process where evidence is accumulated over time toward decision bounds, and a response is made when the evidence reaches a decision boundary. In this task, a possible mechanism for the decision-making process is that the statement presented is compared to pieces of knowledge about the world represented in memory. This comparison results in a weighting where the probability can favor the statement being true or false. Each comparison is counted as a piece of evidence with varying strength for the statement being true or false. Evidence is accumulated over time until the relative evidence weighting is strongly favoring the statement either being true or false (accumulation reaches decision boundary), and a response is made for the favored option. A small bias favoring acceptance of teleological explanations for each piece of evidence increases the chance of accepting a false teleological explanation with more extensive accumulation. Alternatively, the mechanism through which biases arise may be weighting too heavily information that should not influence the outcome of the decision. For example, when evaluating the test statement "The sun makes light so that plants can photosynthesize" the piece of knowledge that plants use light in the photosynthesis process can bias the evaluation of the statement as a whole toward being true, when it is not. This is coherent with evidence showing that low neural gain can broaden attention, which could allow irrelevant information to influence and bias decisions (Eldar et al., 2013, 2021).

The extensive integration account further draws on research showing that the Locus Coeruleus–Norepinephrine system modulates neural gain in the brain which influences neural communication, such that when gain is high, activated neurons become more active, and inhibited neurons become less active (Aston-Jones & Cohen, 2005; Berridge & Waterhouse, 2003; Eldar, Niv, et al., 2016; Joshi et al., 2016). In the sequential sampling process, this means that when gain is high each piece of evidence is more heavily weighted, and fever pieces of evidence are needed to reach a decision boundary (Eldar et al., 2021; Eldar, Niv, et al., 2016). By analyzing trial-by-trial variation in pupil size with the drift-diffusion model, the results strongly support that pupil dynamics reflect changes in neural gain. Larger baseline pupil size was associated with both lower decision threshold and higher drift rate. Thus, larger (tonic) baseline pupil size, indicating higher gain, was associated with less evidence accumulation which led to faster responses and importantly, fewer errors. Conversely, larger phasic pupil dilations were associated with higher decision threshold and lower drift rate. Thus, larger phasic pupil dilations, indicating low neural gain, were associated with more evidence accumulation which led to slower response times and more errors in reasoning. Accordingly, the results corroborate predictions from the extensive integration account.

According to dual-process theories, when Type 2 processes are engaged the normative answer should be more likely. Type 2 processes are indicated by longer response times and more effort, reflected in larger pupil dilations. In this study, we found that normative responses were associated with shorter response times and less effort as reflected in smaller (phasic) pupil dilations, which contradicts dual-process predictions.

Response time in this study was limited but not speeded, that is, time was sufficient as the mean response time was more than two standard deviations below the time limit. The error rate in the test condition in this study was comparable to the error rate in the unspeeded condition in Kelemen et al. (2013). We have no indication of participants having felt time-pressured. But even if so, the speed-accuracy trade-off would have affected the test and control condition similarly (Kelemen et al., 2013).

Pupil dilation leading up to the decision was predicted by response accuracy. On one hand, higher baseline pupil size could indicate an optimal level of arousal and attention (Aston-Jones & Cohen, 2005; Berridge & Waterhouse, 2003). On the other hand, larger pupil dilations could reflect higher uncertainty (Colizoli et al., 2018; Preuschoff et al., 2011; Urai et al., 2017; Yu & Dayan, 2005) on subjectively more difficult trials, where errors indeed are more likely. These explanations are not mutually exclusive but describe separate processes. A less likely explanation, in a dual-process framework, explains the results by rationalization of intuitive errors (however, the authors advise against post-hoc justifications). Additionally, the results could be explained by unsuccessfully invested effort in trials where errors were made. However, there were no differences in effort by condition (see SOM Table S6 for analysis of pupil dilation/effort), which speaks against an explanation of unsuccessfully invested effort. Finding no difference by condition in pupil dilation could be explained by participants not experiencing a difference with regard to conditions in terms of difficulty or not recognizing a need to spend more effort.

Trial-by-trial variation in pupil dilation (and trial baseline and pupil dilation separately) was associated with changes in both threshold and drift rate in the drift-diffusion model, with the model for threshold showing a slightly better fit. This is coherent with findings from Cavanagh et al. (2014) who found that pupil dilation predicted threshold and found a slightly worse fit for drift rate. Other studies have linked pupil dilation to bias and variability in drift rate (de Gee et al., 2020; Leong et al., 2021; Murphy et al., 2014). In this study, pupil dilation had no relation to variability in drift rate. Bias in drift rate was not investigated. The difference in results across studies is probably due to task differences which influence the parameters of the drift-diffusion model, as well as different influences on pupil dilation, that is, arousal, surprise, reward, uncertainty, cognitive effort, and more (Beatty & Lucero-Wagoner, 2000; Laeng et al., 2012). Considering variation in both tasks and influence on pupil dynamics, it is unlikely that pupil dilation would converge on influencing a single parameter of the drift-diffusion model. However, within the context of this study, the influence of both baseline pupil size and pupil dilation on drift rate and threshold fit the predictions from the Extensive integration account.

Feedback-evoked pupil dilations were larger for errors compared to normative responses, which is consistent with an account of pupil dilation signaling uncertainty and surprise (Colizoli et al., 2018; de Gee et al., 2021; Preuschoff et al., 2011; Urai et al., 2017). Additionally, pupil dilations to errors were larger in the control condition indicating higher degree of surprise and higher confidence in the control condition. Higher uncertainty in the test condition compared to the control condition is consistent with the results from drift-diffusion model showing lower drift rate in the test condition indicating higher stimulus difficulty. The results also reflect the behavioral finding of the test condition being more difficult than the control condition.

Individual difference measures of cognitive ability and cognitive motivation were included in the study as predictions from the Default-Interventionist and Smart intuitor accounts differed. Performance on the teleological reasoning task was associated with higher cognitive ability and not cognitive motivation, supporting the Smart intuitor account. The measures of cognitive ability (see SOM for all measures) in this study were included as a convenient indicator of cognitive ability. However, the measures have several limitations and should only be interpreted as indicators of cognitive ability. They should not be interpreted as valid measures of general intelligence. The results should therefore be evaluated with caution. Rational reasoning tasks have been used as a measure dependent on both cognitive ability and cognitive motivation (Stanovich, 2016; Trippas et al., 2015). However, recent evidence suggests performance can be explained by cognitive ability and is not related to cognitive effort (Mækelæ et al., 2023; Otero et al., 2022). We also note that sample size was low and results from individual difference measures should be considered exploratory.

4.1 | Limitations

A limitation of this study is that performance on the teleological reasoning task was not assessed both speeded and unspeeded but with a fixed 4-s time limit for responding. Participants might differ in how time-pressured they felt. Hence, we do not know participants' maximum performance, or how the decision process would unfold without any time restrictions. However, the time to evaluate the truth of statements about the world in real life may not be much longer as there are often implicit time constraints such as flow of conversation, opportunity costs, in addition to cognitive effort costs. Importantly, we do note that there is no known anatomical link between the pupil and the Locus Coeruleus, and the relationship is likely related to common downstream influences (Nieuwenhuis et al., 2011). We therefore have no direct measures of neural gain or the Locus Coeruleus–Norepinephrine system. Variation in pupil size may also be influenced by other factors.

5 | CONCLUSION

Teleological reasoning bias measured as errors in a teleological reasoning task was associated with larger pupil dilations and slower response times. The results support the extensive integration account of bias in reasoning and directly oppose predictions from dual-processing accounts.

AUTHOR CONTRIBUTIONS

Martin Jensen Mækelæ: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; validation; visualization; writing – original draft; writing – review and editing. **Isabel V. Kreis:** Data curation; software; validation; writing – review and editing. **Gerit Pfuhl:** Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision;

validation; visualization; writing – original draft; writing – review and editing.

FUNDING INFORMATION

None.

CONFLICT OF INTEREST STATEMENT

None.

DATA AVAILABILITY STATEMENT

https://osf.io/vk7r4/.

ETHICS STATEMENT

The study was approved by the institutional review board at the Department of Psychology, UiT, The Arctic University of Norway.

ORCID

Martin Jensen Mækelæ https://orcid.org/0000-0002-6791-1218
Isabel V. Kreis https://orcid.org/0000-0002-1022-699X
Gerit Pfuhl https://orcid.org/0000-0002-3271-6447

REFERENCES

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beatty, J., & Lucero-Wagoner, B. (2000). *The pupillary system*. Cambridge University Press.

Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus–noradrenergic system: Modulation of behavioral state and state-dependent cognitive processes. *Brain Research Reviews*, *42*(1), 33–84. https://doi.org/10.1016/S0165-0173(03)00143-7

Busemeyer, J. R., Jessup, R. K., Johnson, J. G., & Townsend, J. T. (2006). Building bridges between neural models and complex decision making behaviour. *Neural Networks: The Official Journal of the International Neural Network Society*, *19*(8), 1047–1058. https://doi.org/10.1016/j.neunet.2006.05.043

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459. https://doi.org/10.1037/0033-295X.100.3.432

Cacioppo, J., Petty, R., Feinstein, J., & Jarvis, B. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*, 197–253. https://doi.org/10.1037/0033-2909.119.2.197

Cacioppo, J., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*(11), 1462–1467. https://doi.org/10.1038/nn.2925

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and Pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, *143*(4), 1476–1488. https://doi.org/10.1037/a0035813

Colizoli, O., de Gee, J. W., Urai, A. E., & Donner, T. H. (2018). Task-evoked pupil responses reflect internal belief states. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-31985-3

de Berker A. O., Rutledge R. B., Mathys C, Marshall L, Cross G. F., Dolan R. J., Bestmann S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*. *7*(1):10996. https://doi.org/10.1038/ncomms10996

de Gee, J. W., Correa, C. M. C., Weaver, M., Donner, T. H., & van Gaal, S. (2021). Pupil dilation and the slow wave ERP reflect surprise about choice outcome resulting from intrinsic variability in decision confidence. *Cerebral Cortex*, *31*(7), 3565–3578. https://doi.org/10.1093/cercor/bhab032

de Gee, J. W., Tsetsos, K., Schwabe, L., Urai, A. E., McCormick, D., McGinley, M. J., & Donner, T. H. (2020). Pupil-linked phasic arousal predicts a reduction of choice bias across species and decision domains. *eLife*, *9*, e54014. https://doi.org/10.7554/eLife.54014

DiYanni, C., & Kelemen, D. (2005). Time to get a new mountain? The role of function in children's conceptions of natural kinds. *Cognition*, *97*(3), 327–335. https://doi.org/10.1016/j.cognition.2004.10.002

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. https://doi.org/10.1038/nn.3428

Eldar, E., Cohen, J. D., & Niv, Y. (2016). Amplified selectivity in cognitive processing implements the neural gain model of norepinephrine function. *Behavioral and Brain Sciences*, *39*, e206. https://doi.org/10.1017/S0140525X15001776

Eldar, E., Felso, V., Cohen, J. D., & Niv, Y. (2021). A pupillary index of susceptibility to decision biases. *Nature Human Behaviour*, *5*(5), 653–662. https://doi.org/10.1038/s41562-020-01006-3

Eldar, E., Niv, Y., & Cohen, J. D. (2016). Do you see the Forest or the tree? Neural gain and breadth versus focus in perceptual processing. *Psychological Science*, *27*(12), 1632–1643. https://doi.org/10.1177/0956797616665578

Epstein S. (1994) Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*. *49*(8):709–24. https://doi.org/10.1037/0003-066X.49.8.709

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239–260. https://doi.org/10.1037/0033-295X.90.3.239

Frederick S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*. *19*(4):25–42. https://doi.org/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: A Journal of the American Psychological Society*, *8*(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. *52*, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models* (Version 0.4.6) [R package]. http://florianhartig.github.io/DHARMa/

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*, *89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Sciences*, *3*(12), 461–468. https://doi.org/10.1016/S1364-6613(99)01402-3

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143. https://doi.org/10.1016/j.cognition.2009.01.001

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*(4), 1074–1083. https://doi.org/10.1037/a0030399

Kreis I., Zhang L., Mittner M., Syla L., Lamm C., Pfuhl G. (2023). Aberrant uncertainty processing is linked to psychotic-like experiences, autistic traits, and is reflected in pupil dilation during probabilistic learning. *Cognitive, Affective, & Behavioral Neuroscience*. *23*(3):905–19. https://doi.org/10.3758/s13415-023-01088-2

Koehler, D. J., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition*, *38*(6), 667–676. https://doi.org/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*(4), 665–682. https://doi.org/10.1037/a0020198

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13852–13857. https://doi.org/10.1073/pnas.1101328108

Kuhn, M. (2015). caret: Classification and regression training. *Astrophysics Source Code Library*, ascl:1505.003.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

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. https://doi.org/10.1037/0003-066X.43.6.431

Leong, Y. C., Dziembaj, R., & D'Esposito, M. (2021). Pupil-linked arousal biases evidence accumulation toward desirable percepts during perceptual decision-making. *Psychological Science*, *32*(9), 1494–1509. https://doi.org/10.1177/09567976211004547

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence*, *30*(1), 81–108. https://doi.org/10.1016/0004-3702(86)90068-8

Mækelæ, M. J., Klevjer, K., Westbrook, A., Eby, N. S., Eriksen, R., & Pfuhl, G. (2023). Is it cognitive effort you measure? Comparing three task paradigms to the need for cognition scale. *PLoS One*, *18*(8), e0290177.

Mækelæ, M. J., & Pfuhl, G. (2019). Deliberate reasoning is not affected by language. *PLoS One*, *14*(1), e0211428. https://doi.org/10.1371/journal.pone.0211428

Mathot, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, *1*(1), 16. https://doi.org/10.5334/joc.18

Murphy, P. R., Vandekerckhove, J., & Nieuwenhuis, S. (2014). Pupil-linked arousal determines variability in perceptual decision making. *PLoS Computational Biology*, *10*(9), e1003854. https://doi.org/10.1371/journal.pcbi.1003854

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology*, *48*(2), 162–175. https://doi.org/10.1111/j.1469-8986.2010.01057.x

Otero, I., Salgado, J. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence*, *90*, 101614. https://doi.org/10.1016/j.intell.2021.101614

Pan, W., Geng, H., Zhang, L., Fengler, A., Frank, M., Zhang, R., & Chuan-Peng, H. (2022, November 1). A Hitchhiker's Guide to

Bayesian Hierarchical Drift-Diffusion Modeling with docker-HDDM. https://doi.org/10.31234/osf.io/6uzga

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian stochastic modelling in python. *Journal of Statistical Software*, *35*(4), 1–81.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, *24*(6), 425–432. https://doi.org/10.1177/0963721415604610

Preuschoff, K., 't Hart, B., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for Noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*, 115.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, 2012.

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, *14*, 170–178.

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*(1), 13289. https://doi.org/10.1038/ncomms13289

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Smullyan, R. M. (1978). *What is the name of this book?: The riddle of Dracula and other logical puzzles*. Prentice-Hall.

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

Stanovich, K. E. (2009b). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (1st ed., pp. 55–88). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K. E. (2016). The comprehensive assessment of rational thinking. *Educational Psychologist*, *51*, 23–34. https://doi.org/10.1080/00461520.2015.1125787

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

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

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*, 672–695. https://doi.org/10.1037/0022-3514.94.4.672

Teigen, K. H., & Keren, G. (2007). Waiting for the bus: When base-rates refuse to be neglected. *Cognition*, *103*(3), 337–357. https://doi.org/10.1016/j.cognition.2006.03.007

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General*, *147*, 945–961. https://doi.org/10.1037/xge0000457

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Toplak, M. E., Liu, E., Macpherson, R., Toneatto, T., & Stanovich, K. E. (2007). The reasoning skills and thinking dispositions of problem gamblers: A dual-process taxonomy. *Journal of Behavioral Decision Making*, *20*, 103–124. https://doi.org/10.1002/bdm.544

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*(7), 1275–1289. https://doi.org/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*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Trippas, D., Pennycook, G., Verde, M. F., & Handley, S. J. (2015). Better but still biased: Analytic cognitive style and belief bias. *Thinking & Reasoning*, *21*(4), 431–445. https://doi.org/10.1080/13546783.2015.1016450

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*(1), 14637. https://doi.org/10.1038/ncomms14637

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*, 757–769. https://doi.org/10.1037/0033-295X.111.3.757

Usher, M., Tsetsos, K., Lagnado, D., & Yu, E. (2013). Dynamics of decision-making: From evidence accumulation to preference and belief. *Frontiers in Psychology*, *4*, 785.

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

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. https://doi.org/10.1037/a0012842

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*, 14. https://doi.org/10.3389/fninf.2013.00014

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Data S1: Supporting Information.

How to cite this article: Mækelæ, M. J., Kreis, I. V., & Pfuhl, G. (2024). Teleological reasoning bias is predicted by pupil dynamics: Evidence for the extensive integration account of bias in reasoning. *Psychophysiology*, *00*, e14532. https://doi.org/10.1111/psyp.14532

Supplementary materials

From article: Teleological reasoning is predicted by pupil dynamics: Evidence for the Extensive Integration Account of Bias in Reasoning.

All reported mixed models from manuscript and pupil waveform.

Analyzed with R version 4.1.2 (2021-11-01)
Packages: lme4, sjPlot
Exclusion criteria applied, N = 39

“” = main text associated with each analysis.

LMM = Linear mixed model
GLMM = Generalized linear mixed model

Response time

“To assess if normative responses were associated with longer (as predicted by dual-process models) or shorter (as predicted by Extensive Integration account) response times, we applied a GLMM with accuracy as outcome (normative – error responses) and z-scored response times and condition as fixed factors and participants as random factors.” – Table S1.

Table S1.
Accuracy ~ 1 + Z-scored Response time + Condition + (1|Participants)

	Accuracy			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	2.44	2.17 – 2.71	17.90	<0.001
Z-scored Response time	-0.48	-0.59 – -0.37	-8.59	<0.001
Condition [test]	-1.18	-1.42 – -0.93	-9.28	<0.001
Random Effects				
σ^2	3.29			
$\tau_{00\ \mathrm{subj\_idx}}$	0.30			
ICC	0.08			
$N_{\mathrm{subj\_idx}}$	39			
Observations	2295			
Marginal R^2 / Conditional R^2	0.154 / 0.225			

Pupil dilation - Decision

“To test if larger or smaller pupil dilations were predictive of correct responses on the teleological reasoning task, we applied a GLMM with accuracy as outcome, PDmax-BL and condition as fixed factors and by-participant random intercepts.” – Table S2.

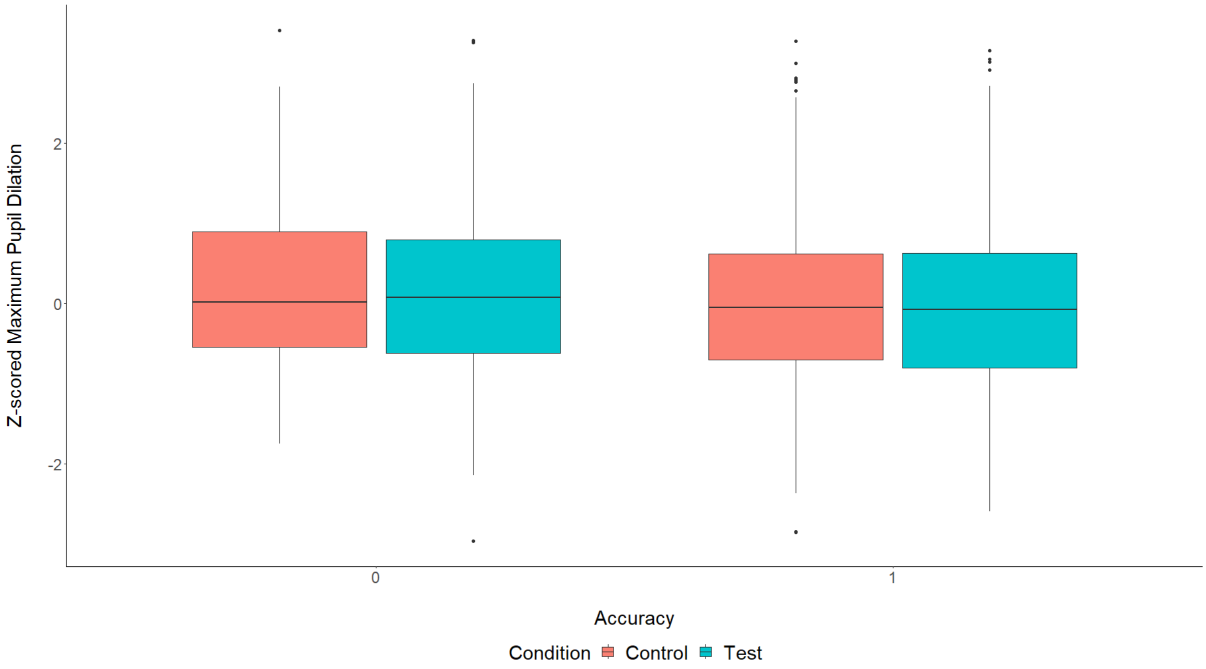
Table S2.

Accuracy ~ 1 + Z- scored PDmax-BL+ Condition + (1|Participants)

	Accuracy			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	2.43	2.17 – 2.69	18.40	<0.001
Z-scored PDmax-BL	-0.19	-0.30 – -0.07	-3.15	0.002
Condition [test]	-1.30	-1.55 – -1.06	-10.51	<0.001
Random Effects				
σ^2	3.29			
τ_{00} subj_idx	0.27			
ICC	0.08			
N subj_idx	39			
Observations	2295			
Marginal R^2 / Conditional R^2	0.112 / 0.180			

Figure S1.

Phasic response (z-scored Maximum Pupil Dilation with Baseline Subtracted) during Listening and Responding.



Note. Phasic response (z-scored maximum pupil dilation with baseline subtracted) during listening and responding in the teleological reasoning task. Responses are separated by condition and accuracy. Accuracy is coded 0 for incorrect responses and 1 for correct responses. Overall, pupil dilation was larger for incorrect trials.

“The Extensive Integration account specifically predicts that lower baseline pupil size and larger pupil dilations are associated with more bias and thus more incorrect responses. To assesses the contribution of both Baseline pupil and PDmax, we applied a GLMM with accuracy as outcome, Baseline pupil, PDmax and condition as fixed factors and by-participant random intercepts.” – Table S3.

Table S3.
Accuracy ~ 1 + Z-scored Baseline pupil + Z-scored Pdmax + Condition + (1|Participants)

	Accuracy			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	2.43	2.17 – 2.69	18.42	<0.001
Z-scored Baseline pupil	0.24	0.08 – 0.39	3.06	0.002
Z-scored PDmax	-0.21	-0.36 – -0.06	-2.76	0.006
Condition [test]	-1.31	-1.55 – -1.06	-10.53	<0.001
Random Effects				
σ^2	3.29			
τ_{00} subj_idx	0.27			
ICC	0.08			
N subj_idx	39			
Observations	2295			
Marginal R^2 / Conditional R^2	0.113 / 0.180			

Individual difference

“To investigate how individual differences in cognitive motivation and cognitive ability influence performance we conducted a linear model with cognitive motivation and cognitive ability as predictors of accuracy in the test condition.” – Table S4

Table S4.
Accuracy ~ Z-scored Rational reasoning + Z-scored Need for Cognition

	Accuracy in test condition		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.74	0.70 – 0.78	<0.001
Z-scored Rational reasoning	0.08	0.04 – 0.12	0.001

Z-scored Need for Cognition	-0.01	-0.05 – 0.03	0.710
Observations	39		
R^2 / R^2 adjusted	0.281 / 0.241		

Pupil – Feedback

"To assess uncertainty and surprise for errors and normative responses in the two conditions we conducted a linear mixed model with Feedback PDmax-BL as outcome and response, condition and their two-way interaction as fixed factors and by-item random intercepts." - Table S5.

Table S5.
Feedback PDmax-BL ~ 1 + Accuracy * Condition + (1|Item)

	Feedback PDmax-BL			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	0.73	0.55 – 0.92	7.72	<0.001
Accuracy	-0.84	-1.04 – -0.65	-8.58	<0.001
Condition [test]	-0.31	-0.54 – -0.09	-2.73	0.006
Accuracy * Condition [test]	0.36	0.11 – 0.60	2.89	0.004
Random Effects				
σ^2	0.92			
$\tau_{00\ \text{trial}}$	0.02			
ICC	0.02			
N_{trial}	75			
Observations	2181			
Marginal R^2 / Conditional R^2	0.052 / 0.069			

We report additional analysis of pupil dilation / effort as mentioned in manuscript.

Pupil dilation – Decision

We applied a LMM to assess if discerning the truth of false teleological statements (test condition) requires more effort than control statements. We compared PDmax-BL in the test condition and the control condition by applying a LMM with PDmax-BL as outcome, and condition and response as fixed factors and by-item random intercepts. – Table S6

Table S6.
Z-scored PDmax-BL ~ 1 + Condition + Accuracy + (1|Item)

	Z-scored Pdmax-BL			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	0.16	0.03 – 0.28	2.49	0.013
Condition [test]	-0.04	-0.12 – 0.05	-0.87	0.384
Accuracy	-0.17	-0.28 – -0.05	-2.90	0.004
Random Effects				
σ^2	0.95			
$\tau_{00\ \mathrm{trial}}$	0.03			
ICC	0.03			
N_{trial}	75			
Observations	2295			
Marginal R^2 / Conditional R^2	0.004 / 0.036			

The results show that there was no significant difference in pupil size due to condition. However, normative answers were significantly associated with smaller pupil dilations. The results show that there was no significant difference in effort, measured as pupil dilation, in the two conditions. However, the results support the EI as smaller pupil dilations are associated with normative answers.

Re-analysis of behavioral data and inclusion of cognitive ability measures

We present a re-analysis of behavioral data with full sample, i.e., exclusions due to pupil data not applied. Additionally, we include two cognitive ability measures for a subsample of participants (N = 33) tested at the beginning of a second test session for a different project.

Additional measures of cognitive ability

The Digit Symbol Substitution Test (DSST) is a timed (90 seconds) paper and pencil measure of processing speed. Participants have to fill in symbols that are paired to each digit (1-9) following a digit-symbol pair code. Performance is measured as the number of symbols correctly coded. The DSST may in addition to processing speed measure psychomotor speed, short-term-visual memory, attention, cognitive flexibility and motivation (Coalson et al., 2010).

The Trail making test (TMT) is a measure dependent on several mental abilities such as psychomotor speed, mental flexibility, visual scanning, and executive function (Halstead-Reitan; Tombaugh, 2004; Salthouse, 2012). The TMT consists of two parts, A and B. In part A (TMT-A) participants are instructed to draw a line between 25 dots, containing numbers from 1 -25, in ascending order. Performance is measured in seconds to complete the task (reverse scored). Part B (TMT-B) consists of 25 dots with both letters and numbers inside. Participants are instructed to draw a line in ascending order, alternating between letters and numbers (1 – A – 2 – B – 3 – C… 13) until the end. Part B scoring is the same as for Part A. We here use time on Part B (reverse scored) as a measure of cognitive ability as Part B has the highest relation to full scale intelligence and fluid intelligence (Corrigan & Hinkeldey, 1987; Salthouse, 2012).

See Table S7 for descriptive statistics for the full sample.

Table S7.
Descriptive statistics

	Mean	SD	Minimum	Maximum
Response time in seconds (n=42)	1.22	0.80	0.01	3.96
Rational reasoning (range 0 to 14) (n=42)	6.93	2.62	2.00	13.00
Need for Cognition (range 18 to 90) (n=42)	54.67	9.98	24.0	74.00
TMT-B in seconds (n = 33)	58.96	19.786	27.75	149.30
DSST score (n = 33)	61.48	11.02	41.00	86.00

Accuracy

A Mann-Whitney U test showed that the percentage of correct responses in the control condition (Mdn = 90.7, SD = 0.6) was significantly higher than the percentage of correct responses in the test condition (Mdn = 73.9, SD = 13.3), U = 1530, p < .001.

Response time

"To assess if normative responses were associated with longer (as predicted by dual-process models) or shorter (as predicted by Extensive Integration account) response times, we applied a GLMM with accuracy as outcome (normative – error responses) and z-scored response times and condition as fixed factors and participants as random factors." – Table S8.

Table S8.
Accuracy ~ 1 + Z-scored Response time + Condition + (1|Participants)

	Accuracy			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	2.37	2.15 – 2.58	21.67	<0.001
Z-scored Response time	-0.53	-0.63 – -0.43	-10.83	<0.001
Condition [test]	-1.12	-1.33 – -0.91	-10.45	<0.001
Random Effects				
σ^2	3.29			

τ_{00} ID	0.18
ICC	0.05
N ID	42
Observations	3053
Marginal R^2 / Conditional R^2	0.162 / 0.206

Individual differences

"To investigate how individual differences in cognitive motivation and cognitive ability influence performance we conducted a linear model with cognitive motivation and cognitive ability as predictors of accuracy in the test condition." – Table S9

Table S9.
Accuracy ~ Z-scored Rational reasoning + Z-scored Need for Cognition

	Accuracy in test condition		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.74	0.71 – 0.77	<0.001
Z-scored Rational reasoning	0.09	0.05 – 0.12	<0.001
Z-scored Need for Cognition	-0.02	-0.05 – 0.02	0.275
Observations	42		
R^2 / R^2 adjusted	0.420 / 0.390		

Additional measures of cognitive ability

A general linear model was applied with accuracy in the test condition as outcome and Rational reasoning, Need for Cognition, Trail Making Test Part B (TMT-B), and Digit Symbol Substitution Test as predictors. – Table S10

Table S10.
Accuracy ~ Z-scored Rational reasoning + Z-scored Need for Cognition + Z-scored Trail Making Test Part B + Z-scored Digit Symbol Substitution Test

	Accuracy in test condition		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.74	0.70 – 0.78	<0.001
Z-scored Rational reasoning	0.07	0.02 – 0.11	0.005
Z-scored Need for Cognition	-0.03	-0.06 – 0.01	0.175
Z-scored Trail Making Test Part B	-0.01	-0.05 – 0.04	0.803

Z-scored Coding	0.04	-0.01 – 0.09	0.077
Observations	33		
R^2 / R^2 adjusted	0.484 / 0.410		

Hierarchical Drift-diffusion model

The models were implemented with the HDDM Python toolbox (Wiecki et al., 2013; HDDM Version 0.9.7, Python Version 3.9 (Patil et al., 2010).

In the first stage we assessed if there were any significant difference due to condition in the main parameters of the model. In the second stage we investigated if trial-by-trial variation in measures of pupil dilation could predict threshold, drift-rate or drift-rate variability. The models in the second stage included either z-scored PDmax-BL “(zpdmaxsubtrbl”) or both z-scored Baseline Pupil (“zbl”) and z-scored PDmax (“zPD”) as predictors.

The second stage only included separate drift-rate for each condition as including both separate threshold and drift-rate only improved DIC by 6 (6196 – 6190 = 6), and the improvement was evaluated as not adding significant improvement to justify the added model complexity. See Table S11 for comparison of all models.

Note that the model with both z-scored Baseline pupil and z-scored PDmax as predictors of threshold had the best fit (DIC 6088). However, the model with pupil dilation with baseline subtracted as a predictor of drift-rate (DIC 6098) had only slightly worse fit (DIC -10).

Arguments

“v” = Drift-rate

“a” = Threshold

“sv” = Drift-rate variability

include=('sv') = include drift-rate variability in model

“Depends_on=” = separate parameter for each condition

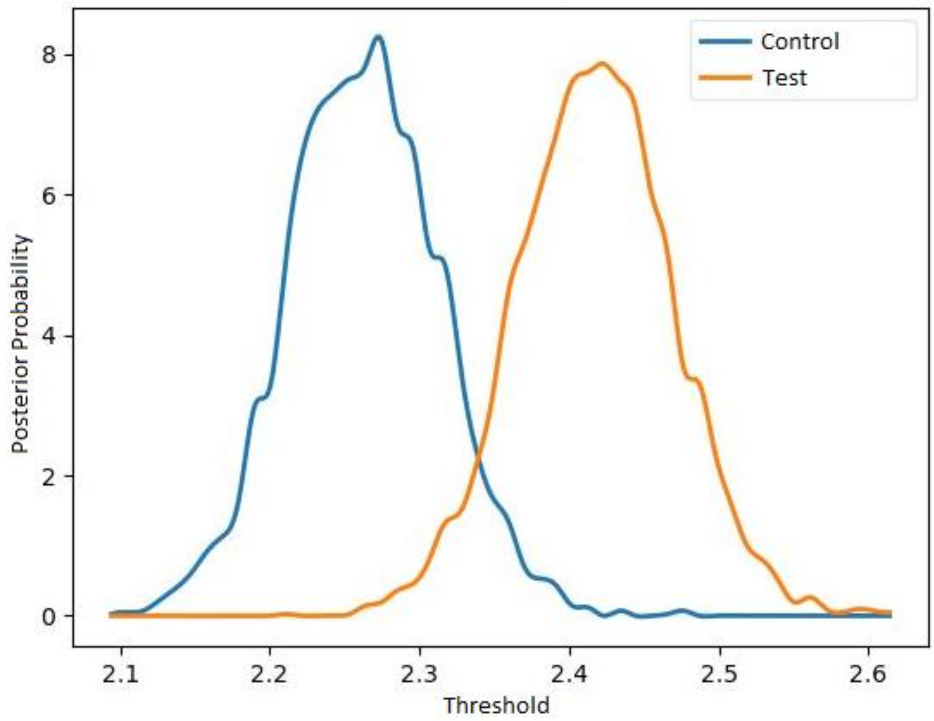
Table S11.
Hierarchical Drift-Diffusion Model Comparison

First stage		
Null model	DIC	6377
include=('sv')	DIC	6370
depends_on={"a": "trialtype"})	DIC	6362
depends_on={"v": "trialtype"}	DIC	6196
depends_on={"v": "trialtype", "a": "trialtype"})	DIC	6190

Second stage		
sv ~ zpdmaxsubtrbl, include=('sv'), depends_on={"v": "trialtype"})	DIC	6140
v ~ zbl+zpdmax, depends on = {"v": "trialtype"}	DIC	6100
v ~ zPD, depends on = {"v": "trialtype"}	DIC	6098
a ~ zPD, depends on = {"v": "trialtype"}	DIC	6093
a ~ zbl+zpdmax, depends on = {"v": "trialtype"}	DIC	6088

Figure S2.

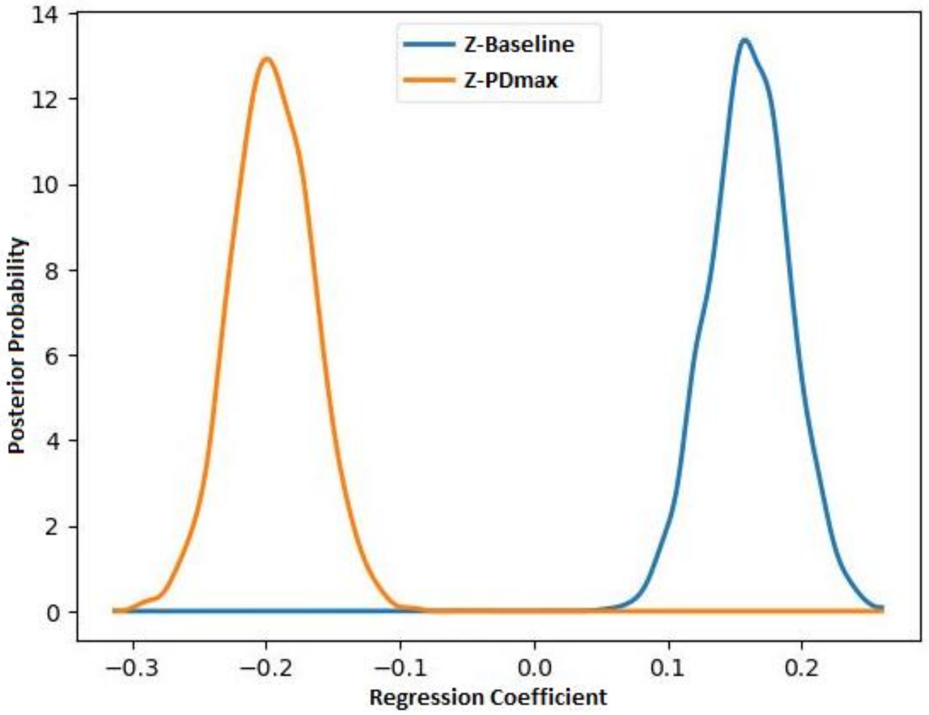
Posterior Estimate of Group Mean Threshold in Test and Control Condition.



Note. Model specification: depends_on={"a": "trialtype"}).

Figure S3.

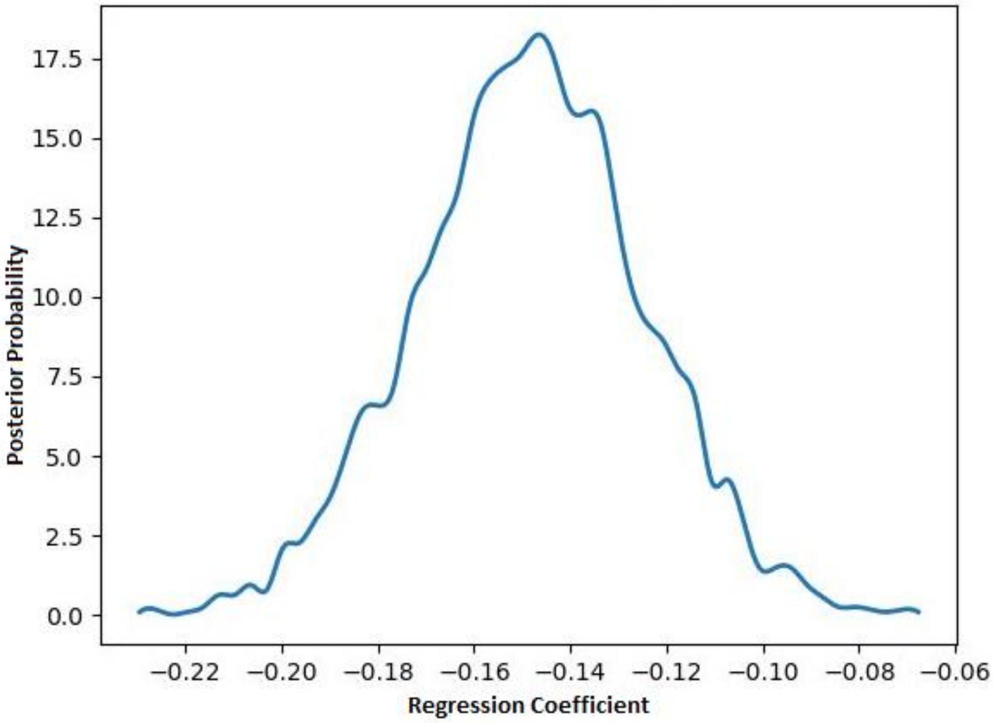
Effect of Z-scored Baseline and Z-scored PDmax on Drift-Rate



Note. Model specification: v ~ zbl+zpdmax, depends_on={"v": "trialtype"}).

Figure S4.

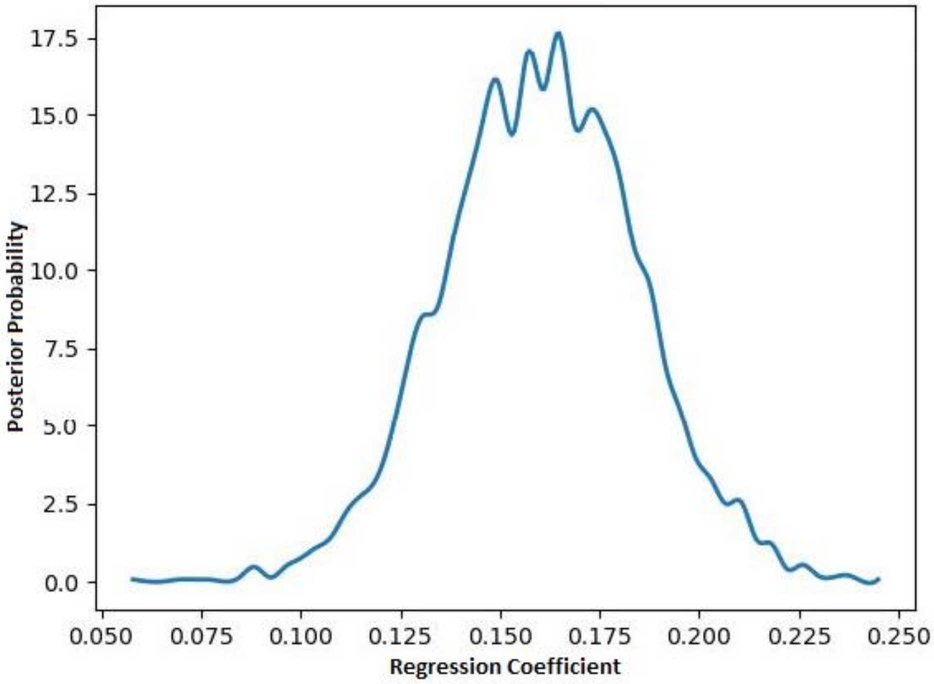
Effect of Z-scored PDmax-BL on Drift-Rate



Note. Model specification: v ~ zpdmaxsubtrbl, depends_on={"v": "trialtype"}.

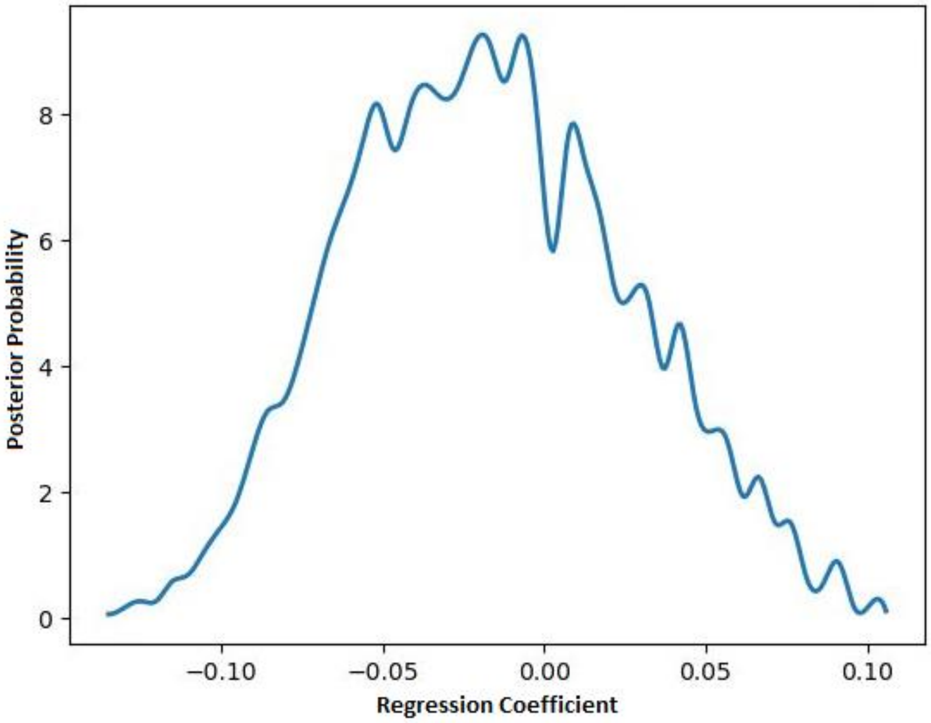
Figure S5.

Effect of Z-scored PDmax-BL on Threshold



Note. Model specification: a ~ zpdmaxsubtrbl, depends_on={"v": "trialtype"}).

Figure S6.
Effect of Z-scored PDmax-BL on Drift-Rate Variability

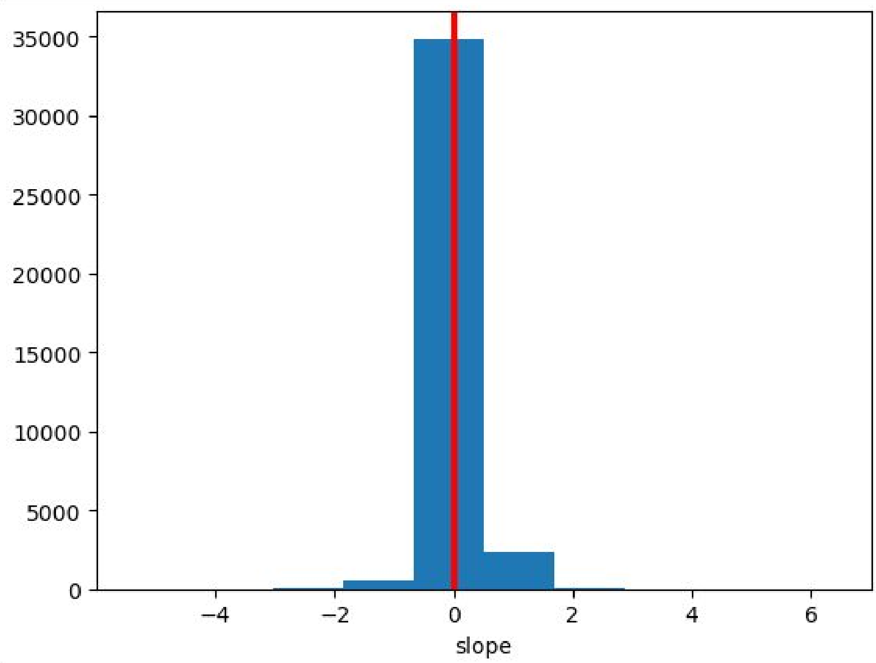


Note. Model specification: sv ~ zpdmaxsubtrbl, include=('sv'), depends_on={"v": "trialtype"}).

Posterior predictive checks for regression

Figure S7.

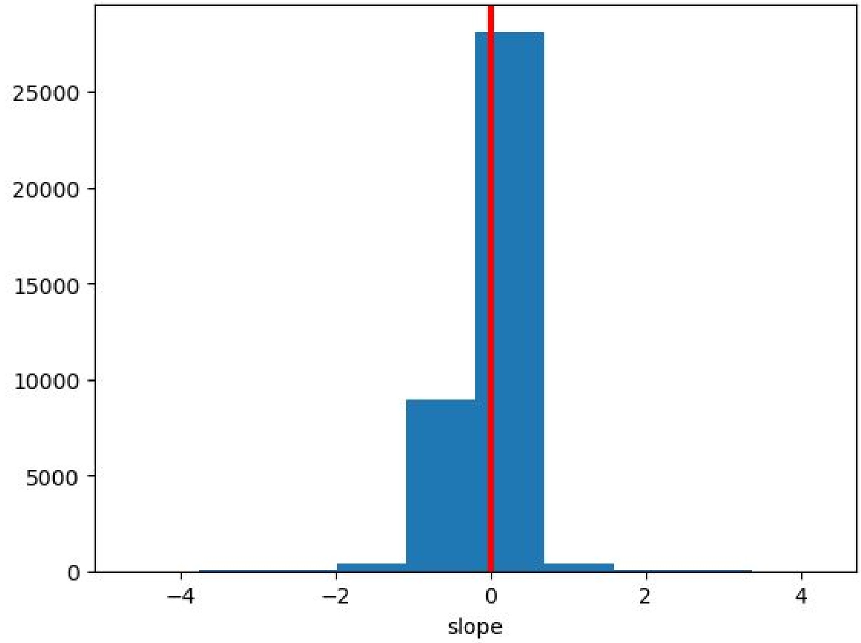
Posterior Predictive for PDmax-BL as a Predictor of Threshold.



Note. Model specification: a ~ zpdmaxsubtrbl, depends_on("v": "trialtype").

Figure S8.

Posterior Predictive for PDmax-BL as a Predictor of Drift-Rate.



Note: Model specification: v ~ zpdmaxsubtrbl, depends_on("v": "trialtype").

Paper 3

The influence of visual attention and cognitive effort on base-rate neglect

Authors: Martin Jensen Mækelæ[1], Isabel V. Kreis[1,2], Gerit Pfuhl[1,3]

Affiliations:

[1]Department of Psychology, UiT The Arctic University of Norway,
[2]Institute of Clinical Medicine, University of Oslo,
[3]Department of Psychology, Norwegian University of Science and Technology,

Electronic addresses:

Martin Jensen Mækelæ: martin.j.makela@uit.no
Isabel V. Kreis: i.v.kreis@medisin.uio.no
Gerit Pfuhl: gerit.pfuhl@ntnu.no

Abstract

Research on errors in reasoning have been crucial in the development of dual-process theories of reasoning. As accumulating research contradicts predictions from classic dual-process models Pennycook et al. (2015) developed a three-stage model of analytic engagement. To test this model, we adapted a base-rate neglect task to be suitable for eye-tracking and pupillometry. The task commonly has a congruent and an incongruent condition, and the base rate information can be given before or after the stereotype information. We alternated the order of information presentation between the two tasks and included a neutral condition where base-rates were equal (non-informative). There were two groups of responders in the task. Stereotype responders mostly neglected base-rate information, responded with the stereotype-congruent response and were insensitive to changes in base-rates. Base-rate responders integrated both base-rate and stereotype information, primarily responded with the base-rate-congruent option and were sensitive to changes in base-rates. Response times were dependent on task structure and usual response. Drift-diffusion modelling yielded no increased decision threshold for conflicting information, disfavoring increased information sampling, response caution or "deliberation". Starting point bias was dependent on task structure. Pupil dilations were related to changing responses from stereotype congruent to base-rate congruent, implicating the Locus Coeruleus – Noradrenaline system in conflict detection and cognitive decoupling. The results question the need for dual-processes.

Introduction

Research on human error in judgement and reasoning has given rise to competing theories of decision-making, where dual-process theories have gained attention and influence in explaining these errors (J. St. B. T. Evans, 2008; Kahneman, 2011). However, a growing body of work contradicts predictions from classical dual-process theories (Bago & De Neys, 2017, 2019; Newman et al., 2017; Raoelison et al., 2020; Raoelison & De Neys, 2019; Thompson et al., 2011), giving rise to a new generation of dual process-models (De Neys & Pennycook, 2019; Pennycook et al., 2015; Raoelison et al., 2020). In this study we test predictions from Pennycook et al. (2015)'s 'Three-stage model of analytic engagement'. By adapting the base-rate task presented in Pennycook et al. (2015) to be compatible with eye-tracking and pupillometry we investigate the role of visual attention and cognitive effort in decision-making. We also model the responses with a drift-diffusion model (DDM).

A dual-process account of errors in reasoning

Errors in reasoning have traditionally been explained in a dual-process framework (J. St. B. T. Evans, 2008; Kahneman, 2011; Tversky & Kahneman, 1973). Dual-process theories broadly propose that decision making relies on two types of processing. Type 1 (intuitive) is automatic, fast, and effortless. Type 2 (deliberate) depends on working memory and is slow and effortful. In the influential default-interventionist (DI) account, errors in reasoning are believed to arise due to overreliance on fast intuitive heuristics and a lack of engagement in deliberate reasoning. The DI account can be exemplified with a classical decision-making task where most participants make a normative error in reasoning by neglecting base-rates. Consider this example problem from De Neys & Gucimic (2008), adapted from Kahneman & Tversky (1973).

In a study 1000 people were tested. Among the participants there were 5 engineers and 995 lawyers. Jack is a randomly chosen participant of this study. Jack is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs.

What is most likely?

a. Jack is an engineer
b. Jack is a lawyer

Participants receive two important pieces of information, the base-rate of each class (5 engineers and 995 lawyers) and a character description (Jack). These pieces of information are assumed to create conflicting outputs from Type 1 and Type 2 processing. According to the DI account, people intuitively process that Jack is a stereotypical representative of an engineer and intuitively, with little effort, answer that Jack is more likely an engineer. On the other hand, integrating the base-rate information with the character description should produce the normative answer, 'Jack is more likely a lawyer' (considered normative as the base-rates are extremely in favor of Jack being a lawyer, but see Gigerenzer (1991, 1994, 1996)). This Type 2 process demands more computational power, more cognitive effort and responding is slower. As humans tend to avoid effort, other things being equal (Kool et al., 2010; Simon, 1957), many participants don't notice the conflict, ignore the base-rates, and give the intuitive incorrect answer (J. St. B. T. Evans, 2003; Kahneman & Frederick, 2002). The default-interventionist account proposes that errors in reasoning occur due to a failure of engaging in deliberate Type 2 processing to produce normative responses and overriding faulty intuitions.

Advances in dual-process models

A growing body of work contradicts predictions from the default-interventionist account. First, even when giving the stereotype congruent response, participants seem to have detected a conflict (De Neys & Glumicic, 2008; Neys et al., 2008; Vartanian et al., 2018). Second, intuitions can be logical, probabilistic and the normative response is often given intuitively (Bago & De Neys, 2017, 2019; Newman et al., 2017; Raoelison et al., 2020). Third, dual-process models have been unclear on how and when Type 2 processing will be engaged. This has given rise to a new generation of dual-process models (De Neys & Pennycook, 2019; Pennycook et al., 2015; Raoelison et al., 2020). Pennycook et al. (2015) proposed a three-stage model of analytic engagement which addresses all three of these critiques (De Neys & Pennycook, 2019; see also Pennycook et al., 2015). In this model many types of intuitions can be produced as a Type 1 process, and competing intuitions are the mechanism causing deliberation. The model suggests that failure in conflict monitoring is an early source of bias, as this leads to the initial response being given (similar to the DI account). The model proposes that if the conflict is detected a Type 2 process will be engaged. This may be rationalizing the initial response, a late source of bias, or it may be cognitive decoupling, leading to suppression and overriding of the initial incorrect response in favor of another response which upon reflection is evaluated as a better response. The separation of early and late sources of bias was demonstrated by adapting the base-rate task for measuring response times (Pennycook et al., 2015).

We adapted the base-rate task to be compatible with eye-tracking and pupillometry to evaluate propositions from the three-stage model of analytic engagement and the role of attention and effort in base-rate neglect. We added a neutral condition where base-rates are equal, thereby one information type does not generate an initial response, assessing the boundary conditions of the model. We model responses with a drift-diffusion model (DDM) as a tool to decompose response time distributions, rather than rely on mean response times (N. J. Evans & Wagenmakers, 2020; Krajbich et al., 2015). Lin et al. (2023) proposed that increased decision threshold in the DDM corresponds with increased deliberation or Type 2 processing. Thus, we investigated whether conflicting information leads to higher decision threshold compared to non-conflicting information.

Using gaze to infer attention and pupil dilation to infer cognitive effort

The use of base-rates depends on the format and structure of the information in the problem and the attention allocated to that information (Barbey & Sloman, 2007; Bar-Hillel, 1980; De Neys & Glumicic, 2008; Gigerenzer et al., 1988; Koehler, 1996; Pennycook & Thompson, 2012). Gaze is a common measure of visual attention in cognitive science as attention and eye movements are interlinked and gaze is known to influence decisions (Armstrong & Olatunji, 2012; Glimcher, 2003; Krajbich, 2019; Petersen & Posner, 2012; Smith & Ratcliff, 2004; Vehlen et al., 2021).

Working memory load and cognitive effort can be gauged with pupillometry (Hess & Polt, 1964; Kahneman & Beatty, 1966; van der Wel & van Steenbergen, 2018). Pupil dilation reflects changes in the Locus Coeruleus (LC) – Norepinephrine (NE) system (Aston-Jones & Cohen, 2005; Eldar, Niv, et al., 2016; Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; Joshi et al., 2016; Joshi & Gold, 2020; Reimer et al., 2016) which might be involved in conflict detection and possibly cognitive decoupling (Arnsten, 2011; Botvinick et al., 2001, 2004; Joshi & Gold, 2020; Shenhav et al., 2013; Spencer & Berridge, 2019; Unsworth & Robison, 2017; Usher et al., 1999). Phasic LC activity has been proposed as a neural interrupt or a network reset signal (Bouret & Sara, 2005; Dayan & Yu, 2006), a putative mechanism for conflict detection.

In summary, we evaluate propositions from the three-stage model of analytic engagement (Pennycook et al., 2015) by adapting a version to measure gaze, investigating the role of attention in the task and adapting a version for pupillometry, investigating cognitive effort and indirectly LC activity. Lastly, we model responses on both tasks with a drift-diffusion model. We expect slower response times in the incongruent condition, gaze location being related to the decision made, and larger pupil dilation when cognitive decoupling occurred.

Methods

Participants

In total 60 participants took part in the study. Participants reported not having any neurological disorder or history of brain disease or surgery, and not taking any drugs or medications affecting the central nervous system. Participants self-rated their English proficiency on a seven-point Likert scale (1 = "understand a few words" and 7 = "master it like native language"). Only participants with a score of four or higher were included as all material was presented in English (Mækelæ & Pfuhl, 2019). The institutional review board at the Department of Psychology at UiT – The Arctic University of Norway approved the study. All participants gave written consent before participating in the study. Participants received either a voucher worth 400NOK (two test sessions) or a voucher worth 150NOK (single test session).

Materials

Base rate task

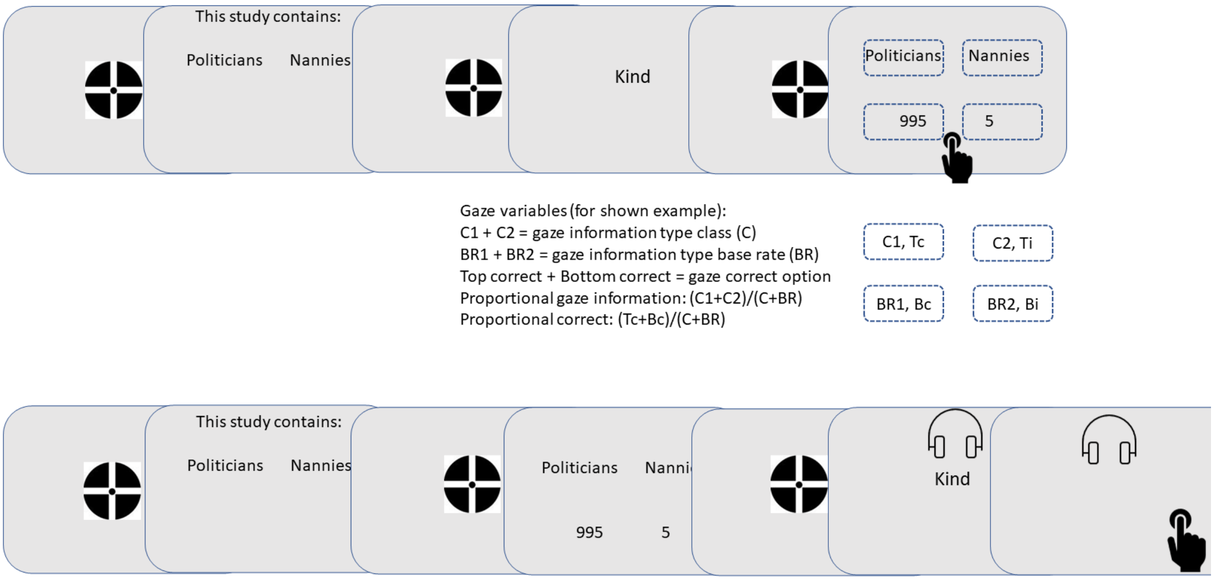
The base-rate task from Pennycook et al. (2015) was used as a template for the two versions of the task presented below. We received the original stimulus materials (list of classes, personality trait, and base-rates) from Gordon Pennycook via personal correspondence and created the adapted versions of the task. Briefly, participants are provided with two pieces of information. 1) Base-rate information. The number of people that are in each class. The base-rates were extremely favoring one class (995 vs. 5, 996 vs. 4, and 997 vs. 3) or they were neutral (e.g., 500 vs. 500). 2) Personality trait. One word describing the person, always fitting to the stereotype of one of the classes and not the other. The task has three conditions; Congruent condition (20 trials), base-rate and personality trait information are favoring the same class. Correct answer is the class favored by both types of information. Incongruent condition (40 trials), base-rate and personality trait information are favoring opposite classes. Correct answer is the class favored by the base-rate information. Neutral condition (20 trials), base-rates are neutral and the personality trait is favoring one class. Correct answer is the class favored by the personality information. The neutral condition creates no conflict but requires decoupling from the base-rate.

Participants first received, on paper, the same background information as in Pennycook et al. (2015, see SOM). Next, participants clicked through a PowerPoint presentation explaining the structure of the task, i.e., a description of the information that were to be presented on the screen during the task, response buttons, and time limit. Participants performed three practice trials, before they completed 80 test trials. There were two versions, counterbalanced across participants. In the gaze version (Fig 1a) the personality trait was given before the base rate information. In the pupillometry version (Fig 2b), the personality trait information was given auditorily after the base rate information. The first 3 seconds are identical.

Figure 1

Trial structure of the two versions of the base-rate task

Top row: Temporal structure of the gaze version. Bottom row: Temporal structure of the pupillometry version.


This study contains:
Politicians
Nannies
Kind
Politicians
Nannies
995
5
Gaze variables (for shown example):
C1 + C2 = gaze information type class (C)
BR1 + BR2 = gaze information type base rate (BR)
Top correct + Bottom correct = gaze correct option
Proportional gaze information: (C1+C2)/(C+BR)
Proportional correct: (Tc+Bc)/(C+BR)
C1, Tc
C2, Ti
BR1, Bc
BR2, Bi
This study contains:
Politicians
Nannies
Politicians
Nanni
995
5
Kind

Legend: A) The trial started by pressing the spacebar followed by a 500ms blank screen (not shown), then by a 500ms fixation cross. For 1800ms the class information was shown, followed by a 200ms fixation cross. Next, for 1800ms the personality trait was shown, followed by a 200ms fixation cross. Up to 4 seconds, the class and base rate information was shown, terminating after a response was made. This was followed by a 200ms fixation cross and 1800ms blank screen used to record post response gaze activity (not shown). There were two response options; left and right, indicated by their relative location on screen. Responses were made by pressing "A" for left option or "L" for right option, on a QWERTY keyboard. Participants were instructed to place their respective index fingers on the response buttons (left index finger on "A" for left option and right index finger on "L" for right option) during task performance. The dotted squares mark the Region of Interest for gaze analysis and were not visible on the screen. B) A trial starts by pressing the spacebar. This was followed by a 500ms blank screen (not shown) and followed by a 500ms fixation cross. For 1800ms the class information was shown, followed by a 200ms fixation cross. Next, for 3600ms both class and base-rate information were displayed, followed by a 200ms fixation cross. Then, a blank screen was displayed while a sound file with the attribute information was played (audio file length, 50ms – 125ms), followed by 2000ms of continued blank screen (allowing recording of pupil dilation). Next, a 200ms fixation cross was presented before a question was presented auditorily. The question was separated by two files. File 1 was the statement "is this person more likely a", audio file length was 2400ms. File 2 was one of the two classes, audio file length 600ms – 1700ms. The files were separated by 100ms. An example question was "Is this person more likely a politician?". After the audio files for the question, participants had 4000ms to respond "yes" or "no", by pressing "A" or "L", respectively. Small notes with "yes" and "no" written visibly were placed behind the respective keys on the keyboard to avoid confusion or working memory demand.

Details regarding the individual difference measures are reported in the SOM.

Procedure

Participants were recruited via flyers at UiT, The Arctic University of Norway. The testing session lasted between 70 – 100 minutes and was conducted individually for each participant during 2020-2021[1]. The base-rate task versions were presented in a counterbalanced order between participants. The structure of testing sessions was as follows. The experimental session started with reading and signing the consent form, and a paper and pencil version of the Trail Making Test (Broshek & Barth, 2000; Salthouse, 2011), followed by the Digit Symbol Substitution Test (Coalson et al., 2010) After instructing on paper and power point on the respective version of the Base-rate task, the participant entered the sound-isolated eye-tracking chamber, and calibration was conducted. Thereafter, participants completed the respective Base-rate task version and rated perceived effort on the N-TLX. Next, participants were offered a 5-minute break and some water. Then participants completed the rational reasoning composite (Mækelæ et al., 2023) and curiosity scale (Kashdan et al., 2018) in Qualtrics (Qualtrics, Provo, UT) on a computer outside the eye-tracking chamber. After completing the individual difference measures, participants were given the instructions (paper and power point) for the other base-rate task version. Participants completed a new calibration procedure on the eye-tracker before starting the base-rate task. They filled out a N-TLX and were debriefed. Participants that took part in the two-day testing, filled out the Need For Cognition (Cacioppo et al., 1984) and personality BFI-20 (Engvik & Clausen, 2011) on day 1. Participants recruited solely for single-day testing filled out the Need For Cognition on a separate day but not the BFI-20.

Gaze and pupil recording

Gaze and pupil size was recorded during both base-rate task versions with a desk-mounted Eyelink 1000 eye tracker (SR-Research, Ontario, Canada) with a sampling rate of 500 Hz. Head position was stabilized by a chinrest at a distance of 65 cm from top of the screen and 69 cm from the screen bottom. Baseline pupil dilation was recorded for two minutes before the start of each task. Participants were instructed to fixate on the center of the screen. Gaze was defined as recorded gaze inside four pre-defined areas of interest (AOI), for both tasks. The areas of interest were four quadrants surrounding the area where class and base-rate information was displayed on screen (Figure 1, AOI not to scale) with upper left / right for class 1 / class 2, and lower left / right for base-rate 1 / base-rate 2. For the Base-rate – gaze version, total gaze time in each AOI was recorded during responding. For the Base-rate – pupillometry version, total gaze time in each AOI was recorded during presentation of the class and base rate information.

Data Processing

Processing of pupil measures were performed in the statistical environment R (version 4.1.2. (R Core Team, 2021). Artifacts such as eyeblinks, rapid changes in pupil size caused by head movements, and alike, were detected based on the signal's velocity (Mathot, 2018). The signal was corrected using linear interpolation. Thresholds for the interpolation window

[1] Due to the pandemic and a 6-month closure of the lab, a fair number of participants could not be recruited (e.g., graduated and left town) for the second testing session, hence we recruited participants solely for the base rate task.

were adapted for each participant, this due to inter-individual differences in signal recovery speed. The interpolated signal was smoothed with a low pass Butterworth filter (3Hz). Artifacts spanning more than 1000 consecutive milliseconds were treated as missing. The signal was visually assessed and trials with remaining artifacts were excluded from further analysis when artifacts occurred during time windows of interest (task-baseline, trial-baseline, attribute, and decision). For the task-baseline measurement the signal within the two-minute pre-task recording of pupil size was averaged to calculate a participant specific task-baseline measure. Subsequently, for each trial, a trial-baseline pupil size was calculated as the average signal during the first 200ms following fixation cross onset. Pupil dilation following presentation of the attribute information was recorded from the end of the sound file with attribute information and the following 2000ms (attribute time-window). The maximum pupil dilation recorded in the attribute time-window was extracted and baseline-corrected by subtracting trial-baseline for each corresponding trial. Pupil dilation during decision was recorded from the end of the question sound files and lasting until a response was given, maximum 4000ms. The maximum pupil dilation recorded in the decision time-window was extracted and baseline-corrected by subtracting the trial-baseline. The task-baseline pupil size, trial-baseline pupil size, attribute pupil dilation, and decision pupil dilation measures were treated as missing (NA) if the interpolated signal was missing in more than 50% of the respective time windows.

Data analyses

Modelling of behavioral responses on the two base-rate tasks with the DDM was performed with Python (version 3.9) (Patil et al., 2010). The models were implemented with the HDDM Python toolbox by using a dockerHDDM (Pan et al., 2022; Wiecki et al., 2013). Linear mixed models were analyzed with the lme4 package (Bates et al., 2015) in R. Non-parametric tests were applied when assumptions of parametric tests were not satisfied. Mann-Whitney U tests were used instead of the Students T-test. Kruskal-Wallis tests with Dunn's post hoc comparisons and Holm corrections were applied instead of one-way ANOVA's.

Participants with accuracy rates more than 3 standard deviations below average in the congruent and neutral conditions were excluded in the base-rate tasks. Response times faster than 150ms and slower than 4000ms were excluded for both base-rate task versions.

Gaze data was used for two measures, proportional gaze at information type (class or base-rate) and proportional gaze at correct option. Proportional gaze at information type indicates if participants are gazing more at the class information (AOI top left and AOI top right) or at the base-rate information (AOI bottom left and AOI bottom right). Proportional gaze at information type was calculated as gaze at the top AOI's (class) divided by total gaze time at all AOI's, resulting in a proportional gaze score where scores above 0.5 indicates more gaze time at the class (top AOI's), and scores below 0.5 indicates more gaze time at the base-rate information (bottom AOI's). Gaze at correct option indicates if participants are gazing more at the option (left or right) which corresponds to the correct response, i.e., if left correct then AOI top left + AOI bottom left. Proportional gaze at correct option was calculated as gaze at correct option (top and bottom for correct side) divided by total gaze at AOI's, where scores above 0.5 indicate more gazing at the correct option and scores below 0.5 indicate more gazing at the incorrect option.

The preprocessing of pupil recordings for the Base-rate pupillometry task revealed some data loss in recordings of pupil size. To ensure the quality of the data, only participants with 40% or more valid trials in the congruent and incongruent conditions were retained for further analyses. Pupil measures were analyzed separately for the respective time-windows of interest. Exclusions due to missing pupil data for each trial were based on the presence of valid data in the time windows of interest. Pupil measures, trial-baseline, attribute pupil dilation, decision pupil dilation, were z-scored within participants.

Exploratory analyses

As we found evidence for two groups of responders in both base-rate tasks, based on their accuracy in the incongruent condition, we further analyzed these groups separately in addition to the analyses for the full sample.

Results and discussion

Base-rate – gaze version: Result and discussion

Two participants were excluded in the Base-rate - gaze version due to low accuracy rates in both the congruent and neutral conditions. Additionally, nine participants failed to make any response in the task and were thus excluded. There were 48 participants after exclusions. Descriptive statistics for the Base-rate - gaze version can be found in Table 1.

Table 1
Descriptive statistics for Base-rate – gaze version

	Congruent		Incongruent		Neutral	
	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.98	0.05	0.63	0.39	0.93	0.08
Mean response time in seconds	1.14	0.65	1.26	0.74	1.46	0.79
proportion of gaze at class	0.65	0.31	0.65	0.31	0.69	0.28
proportion of gaze at correct option	0.73	0.26	0.56	0.32	0.62	0.25

Note. Based on n=48

Accuracy rates reveal two distinct groups, stereotype and base-rate responders

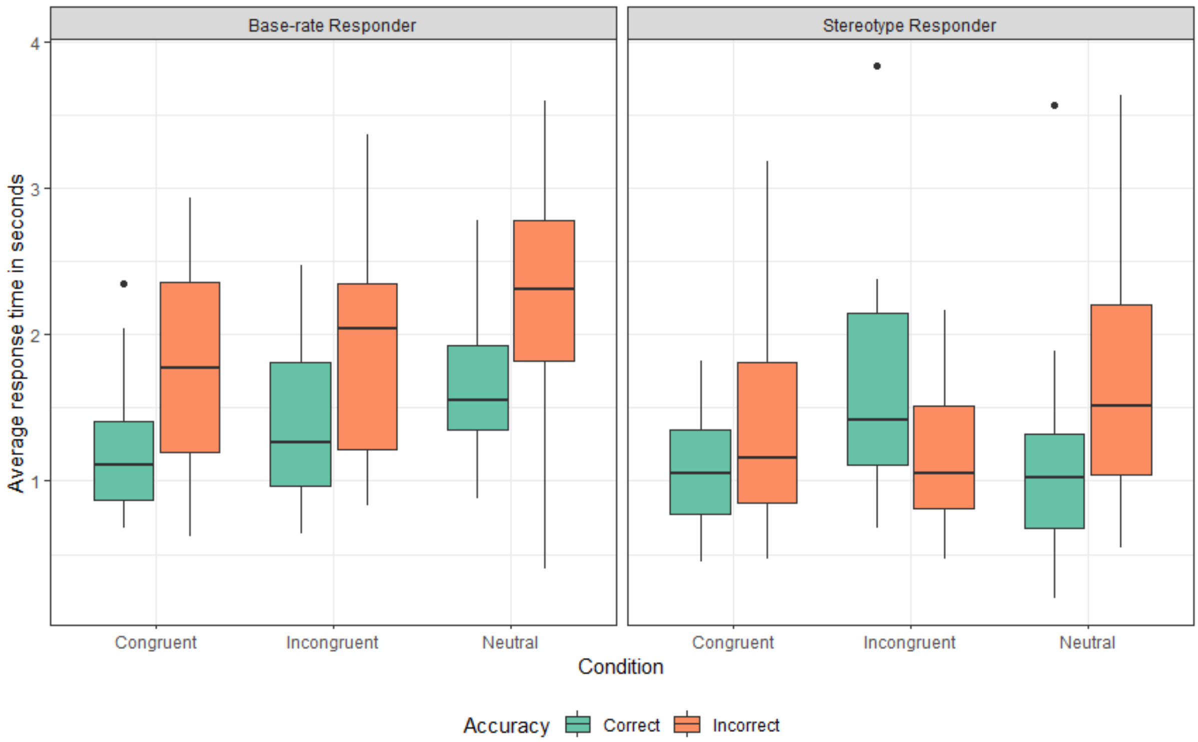
We compared accuracy rates across conditions. A Kruskal-Wallis one-way ANOVA showed a significant group difference, X^2= 27.955, p < .001, df = 2. Pairwise comparisons using Dunn's test showed that there were significant differences in accuracy between a) the congruent and incongruent condition (p < .001), b) the incongruent and neutral condition (p < .001), but not between the congruent and neutral condition (p = 0.346). Consistent with previous studies we find lower accuracy in the incongruent condition compared to the congruent and neutral conditions.

Of particular interest in the base-rate task is the incongruent condition where the attribute information favors another response than the base-rate information. When investigating the distribution of accuracy rates in the incongruent condition across participants we found a bi-modal distribution (Shilling et al., 2002). We classified participants into a) stereotype responders (N = 18) if their accuracy score was lower than 0.5 in the incongruent condition and b) base-rate responders (N = 30) if their score was larger than 0.5. Stereotype responders had an accuracy ranging from 0 to 0.31, with a mean of 0.15 (SD = 0.11). Base-rate responders had an accuracy ranging from .69 to 1, with a mean of 0.91 (SD = 0.12). 18 participants gave the base-rate congruent response every time, and no participant had an accuracy rate between 31% and 69%. Thus, participants tended to choose either stereotype congruent responses or base-rate congruent responses in the incongruent condition. As these responder groups were clearly distinct, we used them for further analysis.

Response times across conditions differ for the base-rate responders but not for the stereotype responders

A common finding in dual-process research and the base-rate task is that conflicting information causes an increase in response times. We expected longer response times in the incongruent condition compared to the congruent condition, and had no specified prediction for the neutral condition. Response times in the congruent, incongruent, and neutral conditions were significantly different, X^2 = 90.585, p < .001, df = 2. Pairwise comparisons using Dunn's test showed that there were significant differences in response times between the congruent and incongruent condition (p < .001), the incongruent and neutral condition (p < .001), and between the congruent and neutral condition (p < .001). Consistent with previous studies the congruent condition had the fastest response times, followed by the incongruent condition. The neutral condition (Table 1, Fig. 2) showed longer response times than both the congruent and incongruent condition.

Figure 2
Response times across conditions for the stereotype and base-rate responders and separate for correct and incorrect responses



When investigating response times across conditions separately for the stereotype responders and the base-rate responders, we found no difference in response times across conditions for the stereotype responders, X^2 = 0.181, p = .913, df = 2. Base-rate responders had significant response time differences across the three conditions, X^2 = 169.386, p < .001, df = 2. Dunn's post hoc comparisons showed that all three conditions were significantly different for the base-rate responders (all p-values < .001). As seen in Figure 2, the base-rate responders had slower response times in the incongruent condition compared to the congruent condition and significantly longer response times in the neutral condition.

Response times in the incongruent condition are dependent on response preference

A common finding supporting the intuitive nature of stereotype processing (Type 1 processing) is that stereotype congruent responses are given faster than base-rate responses in the incongruent condition. A Mann-Whitney U test (full sample) showed that response times for correct (base-rate congruent) responses (*M* = 1.31, *SD* = 0.70) were significantly slower compared to incorrect (stereotype congruent) responses (*M* = 1.16, *SD*= 0.78) in the incongruent condition, U = 295717.500, d = -0.194, p < .001. This was mainly driven by the stereotype responders, i.e., base-rate consistent responses were slower than stereotype consistent responses, U = 14963.500, d = -0.42, p < .001. The opposite was the case for the base-rate responders where base-rate consistent responses were faster than the stereotype congruent responses, U = 70646.500, d = 0.38, p < .001. Participants were faster when responding with their majority response than when giving the response opposite to their typical response (Fig. 2).

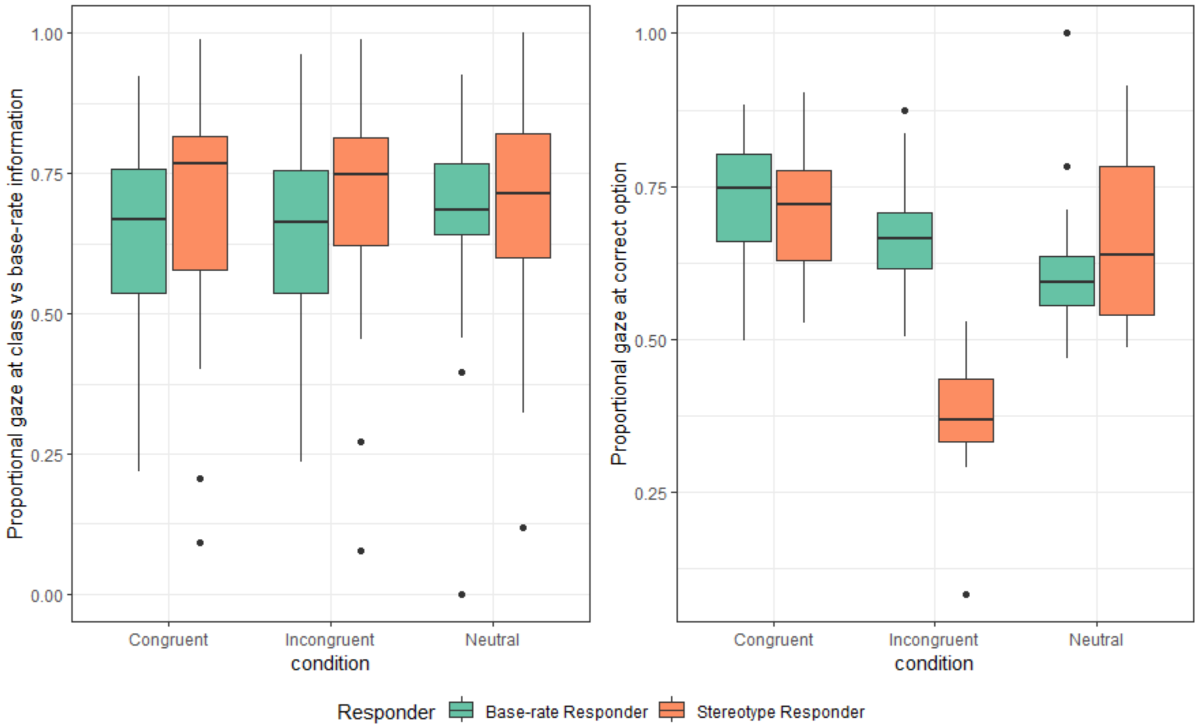
Gaze at information type differ across conditions for the base-rate responders but not for the stereotype responders

Participants tended to look more at the class (stereotype) information compared to the base-rate information (proportional gaze > 0.5) in the congruent (*M* = 0.65, *SD* = 0.31), incongruent (*M* = 0.65, *SD* = 0.31) and neutral conditions (*M* = 0.69, *SD* = 0.28).

Conflicting information influenced participants proportional gaze at information type, X^2 = 7.677, p = .022, df = 2. A Dunn's post hoc pairwise comparison showed no difference between the congruent and incongruent condition (p = 0.922), but in the neutral condition the proportional gaze at information type was significantly higher compared to the congruent (p = 0.04) and the incongruent (p = 0.029) condition. This was driven by the base-rate responders, i.e., there was a significant difference in gaze at information type, X^2 = 14.386, p < .001, df = 2. Dunn's post hoc comparisons showed no difference between the congruent and incongruent condition (p = .960), but in the neutral condition the proportional gaze at information type was significantly higher compared to the congruent (p = 0.003) and the incongruent (p <.001) condition. The stereo-type responders had no difference in gaze at information type across the three conditions, X^2 = .050, p = .975, df = 2.

Figure 3

Proportional gaze at class information and gaze at correct option by condition separate for stereotype and base-rate responders



Legend. Base-rate responders looked significantly less at the class and more at the base-rate information (left-hand panel) compared to the Stereotype responders (U = 1.437, r_{rb}= 0.10, p < .001). This was largely due to differences in gaze in the congruent and incongruent condition. Base-rate responders looked more at the correct option in the incongruent condition (right-hand panel).

Further, we separated correct (base-rate congruent) and incorrect (stereotype-congruent) responses in the incongruent condition. A Mann-Whitey U test showed that correct (base-rate congruent) responses (*M* = 0.63, *SD* = 0.28) were associated with significantly less time looking at class, compared to incorrect (stereotype congruent) responses (*M* = 0.68, *SD* = 0.34, U = 391078.500, d = 0.16, p < .001). Participants giving the base-rate congruent response did look more at the base-rate information (Fig 3, left-hand panel).

Overall, the results show no significant differences in proportional gaze at information type for the congruent and incongruent condition. However, when the base-rates were non-informative (neutral condition) participants looked longer at the stereotype information. Conflicting information between base-rates and attribute (stereotype) does not lead to overall increased investigation of the base-rate information. However, participants spend less time on the base-rate information when the base-rates are non-informative for their decision, i.e., the neutral condition (Fig. 3).

Drift-diffusion modelling - Lower drift-rate not higher threshold in incongruent condition

According to Pennycook et al. (2015) conflicting responses engage Type 2 processing. In the drift-diffusion model an increased threshold would indicate increased information sampling and more cautious responding, and the incongruent condition should lead to a higher threshold (Lin et al., 2023; Pennycook et al., 2015).

The model with the best fit to the data was a model with separate drift-rate for each condition (see SOM for model comparison). Drift-rate was lowest in the incongruent condition (p-values < .001) and highest in the neutral condition (p-values < .016). Separate threshold by condition did not improve model fit. These results indicate that the conflicting information in the incongruent condition does not lead to increased threshold or engage Type 2 reasoning, rather the stimulus difficulty is higher as seen in lower drift-rate.

An additional analysis of response bias revealed a bias towards stereotype congruent responses in the incongruent condition (see SOM).

Discussion Base-rate - gaze version

In the Base-rate – gaze version, the data reproduced behavioral results consistent with previous literature, namely lower accuracy in the incongruent condition and longer response times in the incongruent condition compared to the congruent condition (Pennycook et al., 2014, 2015; Pennycook & Thompson, 2012). Importantly, we found two groups of responders (or response strategies) base-rate responders and stereotype responders. The base-rate responders were affected by base-rates whereas the stereotype responders were not affected. This was reflected in both response times and proportional gaze at information type. Base-rate responders showed slower response times in the incongruent condition compared to the congruent condition, and the slowest response times in the neutral condition. Base-rate responders changed their proportional gaze towards the class information when the base-rates were not informative in the neutral condition. Stereotype responders did not show significant differences in response times or gaze across the three conditions. Thus, base-rate responders integrated both types of information whereas stereotype responders relied on the stereotype information, and largely ignored the base-rate information. This suggests that stereotype responders do not detect conflicting information and “conflict detection failure” is a significant source of “biased responses” in this task.

An important insight gained from analyzing separately the stereotype responders and base-rate responders is that they differed markedly in their response times in the incongruent condition. The classical pattern is that stereotype-congruent responses are faster compared to base-rate-congruent responses, as also found for the full sample. However, when analyzing the two groups separately we found that base-rate responders responded faster when giving the base-rate congruent response than when giving the stereotype-congruent response, whereas the stereotype responders showed the classical pattern. Thus, overriding one’s dominant or default response seems to require more time. This explains also why the base-rate responders but not the stereotype responders responded slow in the neutral condition.

Investigating the influence of gaze on responses showed that participants tended to look more at the class information. This may be due to the ease of processing the relative group difference using extreme base-rates. The judgement of which group is congruent with the stereotype may have been comparatively harder. Interestingly, for base-rate responders the neutral condition was the condition where they looked proportionally the least at the correct option, indicating that they may have found the task of finding the stereotype congruent response harder compared to when they could use the base-rate information for their judgement. Participants tended to look more at the option they ended up choosing. Further, both gaze at information type and option type was associated with responses, with the latter having a larger effect. These results are in line with recent work investigating the effect of gaze on choice (Yang & Krajbich, 2023).

A significant contribution in this work is the modelling of response times with the drift-diffusion model. If conflicting information does trigger Type 2 processing there would be a larger threshold in the incongruent compared to the congruent condition, indicating increased information sampling and response caution (Lin et al., 2023). Contrary to this prediction we found a lower drift-rate in the incongruent condition but not larger threshold. Thus, conflicting information does not trigger increased information sampling, response caution or “deliberation”, rather, task difficulty is higher, leading to a slower drift towards response boundaries, and longer response times.

Base-rate pupillometry version: Result and discussion

Two participants were excluded in the Base-rate Pupillometry version due to low accuracy rates in both the congruent and neutral conditions. There were 55 participants with valid behavioral data after exclusions. The pre-processing of pupil recording revealed some data loss in recordings of pupil size (see “Data pre-processing” section). Descriptive statistics for the Base-rate pupillometry task are presented in Table 2.

Table 2
Descriptive statistics for the Base-rate pupillometry task

	Congruent		Incongruent		Neutral		Valid N
	Mean	SD	Mean	SD	Mean	SD	
Accuracy	0.95	0.06	0.57	0.35	0.90	0.15	55
Response time	1.02	0.34	1.07	0.36	1.00	0.30	55
PS: Trial baseline (BL)	34.16	4.32	34.17	4.40	34.21	4.51	47
PD: Decision	36.30	4.46	36.64	4.55	36.67	4.48	38
PD: Decision - BL	2.25	2.98	2.48	2.74	2.53	2.74	38
PD: Attribute	36.32	4.38	36.54	4.44	36.60	4.38	37
PD: Attribute - BL	2.26	2.68	2.45	2.53	2.58	2.47	37

Note. Abbreviations: PD = pupil dilation, PS = pupil size.

Accuracy ratings identify two response groups

Accuracy was high in the congruent and neutral conditions, and low in the incongruent condition (Table 2). A Kruskal-Wallis one-way ANOVA showed a significant difference, X^2 = 63.622, p < .001, df = 2. Dunn’s post-hoc test yielded significant differences in accuracy between the congruent and incongruent condition (p < .001), and the incongruent and neutral condition (p < .001), but not between the congruent and neutral condition (p = .120). Similar to the Base-rate – gaze version we find lower accuracy in the incongruent condition compared to the congruent and neutral condition.

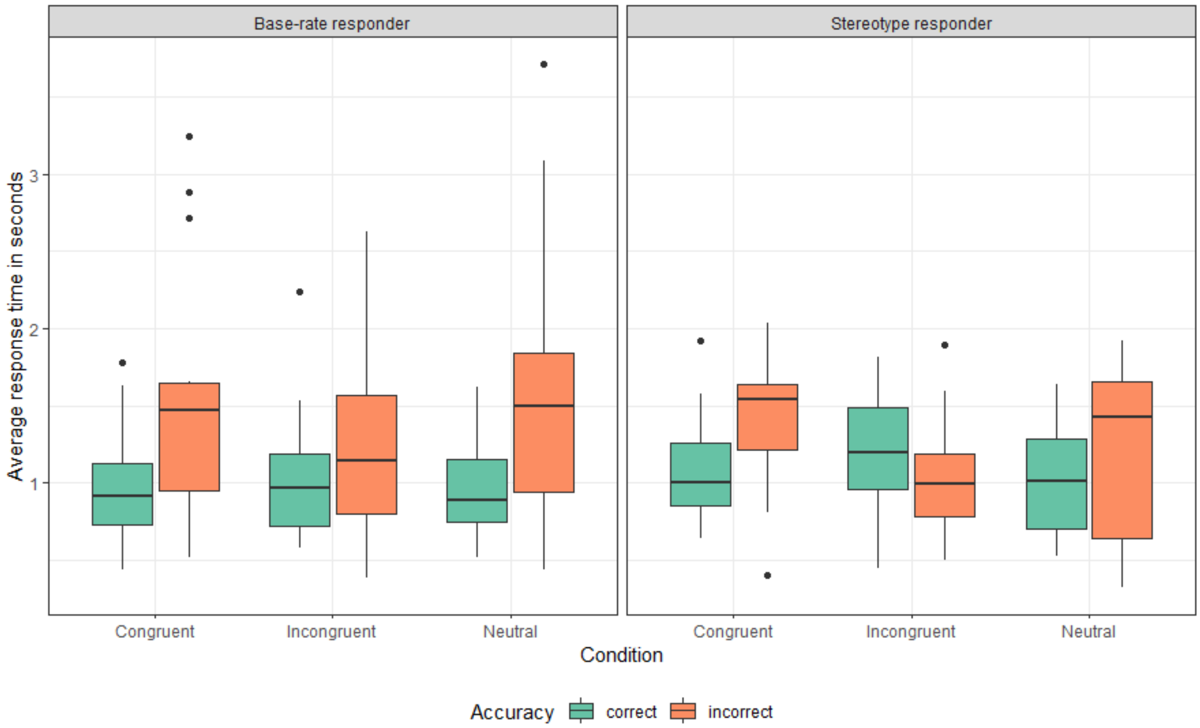
We found strong evidence for a bi-modal distribution with base-rate responders showing substantially higher accuracy rates (N = 34, *M* = 0.83, *SD* = 0.11) compared to stereotype responders (N = 20, *M* = 0.14, *SD* = 0.11). Two participants had changed group membership from stereotype-responder in the gaze version to base-rate responder in the pupillometry version. One participant had an accuracy rating of 0.5, and was excluded from analyses regarding the two groups. Participants gave more base-rate congruent responses in the incongruent condition as the task progressed (see SOM, section learning effects).

Response times differ for base-rate responders but not for stereotype responders

In the full sample, response times did not significantly differ between conditions (X^2 = 5.439, p = .066, df = 2). This was also true for stereotype responders (X^2 = 0.912, p = .634, df = 2). Base rate responders differed in response time by condition (X^2 = 11, p = .004, df = 2). Pairwise comparisons using Dunn’s test showed significant differences in response time between the congruent and incongruent condition (p = .005), but not between the incongruent and neutral condition (p = .083), or between the congruent and neutral condition (p = 0.339). The neutral condition was not significantly different from the other conditions, however consistent with the Base-rate – gaze version and previous literature the incongruent condition was significantly different from the congruent condition. Thus, there was a slowing of response times due to incongruent information for base-rate responders but not for stereotype responders (Fig. 4).

Figure 4

Response times across conditions, separately for correct and incorrect responses and separately for base-rate responders and stereotype responders



Response times in the incongruent condition are dependent on response preference

In the incongruent condition stereotype-congruent responses were significantly slower than base-rate-congruent responses, U = 527214.00, p < .034, d = 0.06. Next, we separately analyzed responses in the incongruent condition for base-rate responders and stereotype responders. Stereotype responders had significantly slower base-rate congruent responses than stereotype congruent responses, U = 26840.500, p = .017, d = -0.15. For base-rate responders stereotype-congruent responses were significantly slower than base-rate congruent responses, U = 130713.000, p < .001, d = 0.19. Thus, we replicate the finding from the Base-rate gaze version that participants are faster when responding with their majority response and slower when giving the response opposite to their majority response (Fig. 4).

Base-rate responders look more at the base-rate information

Similar to the gaze version of the task we find that base-rate responders, but not stereo-type responders look more at the base-rate information (see SOM, section 3. Base-rate - pupillometry version. Fig S1).

Pupil size is related to performance, support for conflict detection and cognitive decoupling

For the pupil analysis we first assessed if there were significant differences in pupil size in the attribute time window. If conflicting information does engage Type 2 reasoning, we expect to see larger pupil dilations following the attribute information in the incongruent condition.

In the attribute time window, there was a significant difference between conditions F(2,111) = 4.344, η2 = 0.07, p = .015. Post hoc comparisons showed that pupil dilations were significantly larger in the neutral condition compared to the congruent condition (p = .011), but there was no significant difference between the congruent and incongruent condition (p = .320), nor between the neutral and incongruent condition (p = .295). When analyzing the two groups of responders separately there were no significant difference in pupil dilation between conditions in the attribute time window for the stereotype responders, F(2,27) = 1.059, η2 = 0.07, p = .361. However, there was a significant difference in pupil dilation between conditions in the attribute time window for the base-rate responders, F(2,78) = 3.279, η2 = 0.08, p = .043. Post hoc comparisons showed that pupil dilations in the neutral condition was significantly larger compared to the congruent condition (p = .034), but there was neither a difference between the incongruent and congruent conditions (p = .546), nor between the incongruent and neutral conditions (p = .299).

Contrary to the predictions of the three-stage model of analytic engagement there was no significant difference in pupil dilation in the attribute time window between the congruent and incongruent conditions. However, there was a significant difference between the congruent and neutral conditions. This effect was significant for the base-rate responders but not for the stereotype responders. A possible explanation for these results is that the base-rate responders have to expend more cognitive effort when they have to change strategy from relying on the base-rate information to using the attribute information to find the correct stereotype.

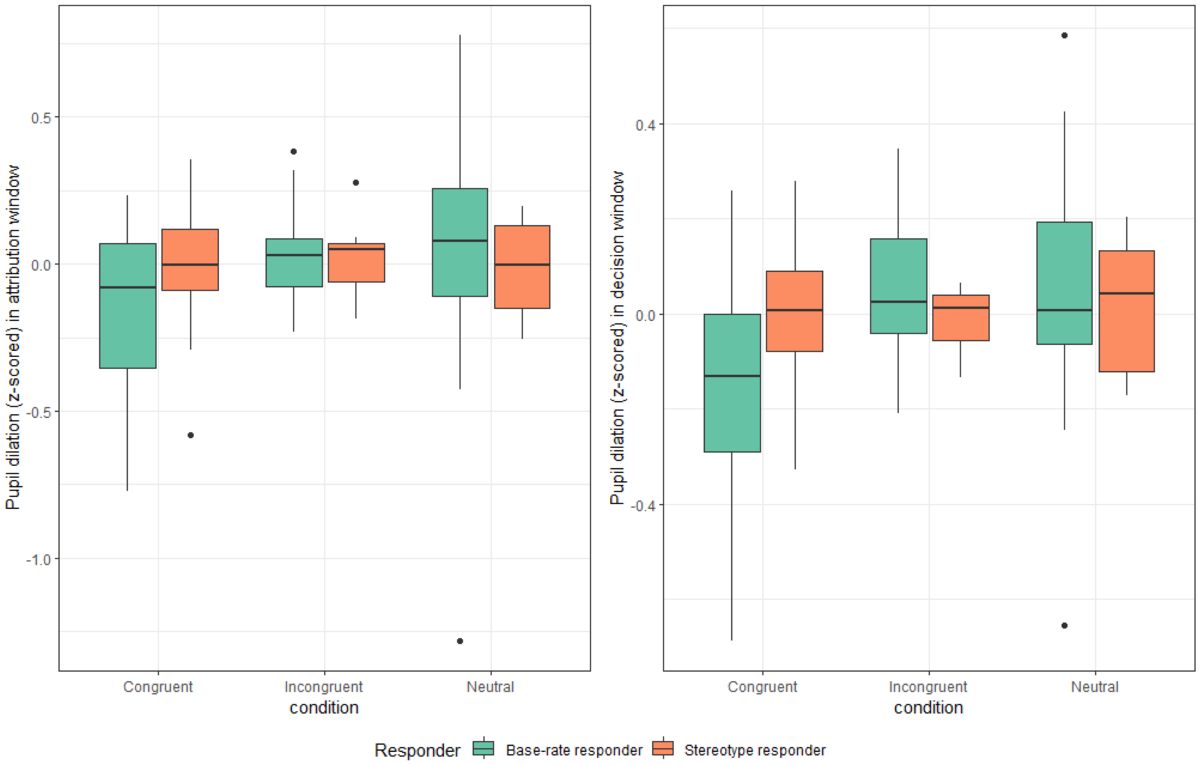
Of particular interest for the three-stage model of analytic engagement are responses in the incongruent condition. The model proposes that conflict detection is necessary for Type 2 processing, which is more likely associated with normative base-rate congruent responses.

A GLMM with response as outcome, pupil dilation in the attribute time window as a fixed factor and subjects as a random factor showed that pupil dilation was not a significant predictor of responses in the incongruent condition (β = 0.15, SE = 0.08, z = 1.81, p = .071). However, when analyzing the stereotype group separately pupil dilation in the attribute time window was a significant predictor of normative responses (β = 0.37, SE = 0.17, z = 2.20, p=

.028). In contrast, for the base-rate group, pupil dilation was not a significant predictor of responses (β = 0.08, SE = 0.10, z = 0.78, p = .433).

These results indicate that conflict detection, measured as pupil dilation following the attribute information was associated with changing responses from the usual response (stereotype congruent) to the normative (base-rate congruent) response for stereotype responders. However, there was no significant effect that could be detected across participants, nor for base-rate responders (Fig 5).

Figure 5
Pupil dilation (z-scored) in the attribute time window and decision time window for the three conditions and separately for base-rate responders and stereotype responders



Next, we compared pupil dilations in the decision time window across conditions. If conflicting information engage type 2 processing there should be larger pupil dilations in the incongruent condition.

The results showed that there was a significant difference in pupil dilations across conditions in the decision time window F(2,111) = 4.004, η2 = 0.07, p = .021. Post hoc comparisons showed that pupil dilations in the incongruent condition was not significantly larger (although descriptively larger) compared to the congruent condition (p = .082). However, pupil dilations in the neutral condition were significantly larger than the congruent condition (p = .024). There was no significant difference between the neutral and incongruent condition (p = .871). When analyzing pupil dilations in the decision time window for the stereotype responders there were no significant difference across conditions F(2,27) = 0.095, η2 = 0.01, p = .910. In contrast, for the base-rate responders there was a significant difference F(2,78) = 4.842, η2 = 0.11, p = .010. Post hoc analyses revealed that pupil dilations in the incongruent condition were significantly larger than the congruent condition (p = .034). Additionally, pupil dilations were larger in the neutral condition compared to the congruent condition (p = .016). There was no significant difference between the incongruent and neutral conditions (p = .958).

The results show that there is an effect of condition on pupil dilations before decisions, restricted to base-rate responders. As proposed by the three-stage model of analytic engagement conflicting information in the incongruent condition was associated with larger pupil dilations compared to the congruent condition, restricted to base-rate responders. Importantly, the largest pupil dilations were seen in the neutral condition, thus indicating that other alternative sources, such as having to change strategy, may also engage Type 2 processing.

To assess if cognitive decoupling was associated with normative responses in the incongruent condition. A GLMM with response as outcome and pupil dilation before decision as fixed factor and participants as random factor was conducted. The result shows that pupil dilations before decisions were not a significant predictor of responses in the incongruent condition (β = 0.13, SE = 0.09, z = 1.51, p = .130). When analyzing the two groups of responders separately pupil dilation was neither a significant predictor of responses in the incongruent condition for stereotype responders (β = 0.31, SE = 0.17, z = 1.88, p = .061), nor for base-rate responders (β = 0., SE = 0.09, z = 1.51, p = .130).

The results indicate that pupil dilations before decisions were not a significant predictor of normative responses in the incongruent condition. However, it should be noted that for the stereotype responders there is an effect in the expected direction.

To test if pupil size is predictive of performance on the Base-rate pupillometry task across conditions we conducted a series of GLMM's with response as outcome, condition and pupil measures (trial-baseline, attribute and decision) as fixed factors and participants as random factor. A model with condition and trial baseline pupil size as predictors of normative responses showed that smaller trial baseline pupil size was a significant predictor of normative responses (β = -0.25, SE = 0.06, z = -4.33, p < .001) and explained 23.5% of the variance.[2] A model with condition and pupil dilation following attribute presentation as predictors showed that pupil dilations following the attribute was a not a significant predictor of normative responses (β = 0.10, SE = 0.06, z = 1.56, p = .119), but explained 23.2% of the variance. A model with condition and pupil dilation leading up to decision as predictors of responses showed that larger pupil dilations were a significant predictor of normative responses (β = 0.15, SE = 0.06, z = 2.37, p = .0018) and explained 22.1% of the variance. There were no significant interactions between pupil measures that could predict responses in any of the models (see SOM). Thus, we find that smaller baseline pupil size and

[2] The effect may be driven by the incongruent condition, see SOM (section 6. The role of the Locus Coeruleus – Noradrenaline system).

larger pupil dilation before decisions are predictive of normative responding across conditions, whereas pupil dilations following attribute presentation was not.

Drift-diffusion modelling - Lower drift-rate not higher threshold in incongruent condition

Model comparison of the behavioral data showed that a model with separate drift-rate by condition provided the best fit. In this model drift-rate in the incongruent condition was significantly lower compared to the congruent (p < .001) and neutral (p < .001) conditions, while there was no significant differences between drift-rates in the congruent and neutral condition (p = .174). An additional analysis of response bias revealed a bias towards base-rate congruent responses in the incongruent condition (see SOM “Hierarchical drift-diffusion modelling” for details). Thus, similar to the gaze version, task difficulty was higher in the incongruent condition, but did not lead to more deliberation (higher decision threshold).

Discussion Base-rate pupillometry version

The results from the Base-rate pupillometry version replicated a number of important findings from the Base-rate - gaze version. We again found that there were two groups of responders, base-rate responders and stereotype responders. There was an increase in response time in the incongruent condition compared to the congruent condition for base-rate responders, however this effect was not significant for stereotype responders or in the full sample. The results show that the base-rate responders looked significantly more at the base-rate information compared to the stereotype responders. Additionally, base-rate responders looked significantly less at the base-rate information in the neutral condition (when the information was non-informative) compared to the congruent and incongruent condition. This effect was not significant (although trending in the same direction) for the stereotype responders. Thus, we again find that base-rate responders seem to be influenced by changes due to condition whereas the stereotype responders do not show significant changes in response times or gaze due to condition.

The results from analyzing pupil size during task performance shows that smaller trial-baseline pupil size and larger pupil dilations before a decision is predictive of normative responses. Thus, we find evidence that attention and cognitive effort is related to performance on the Base-rate pupillometry version. This is consistent with previous work showing a relationship between pupil dilation and performance (Aston-Jones & Cohen, 2005; Laeng et al., 2011; van der Wel & van Steenbergen, 2018; van Steenbergen & Band, 2013).

Pupil dilations following attribute presentation was significantly different across conditions, with larger pupil dilations in the neutral condition compared to the congruent condition. This effect was present for the base-rate responders but not for the stereotype responders. However, when analyzing responses in the incongruent condition stereotype responders showed larger pupil dilations after attribute presentation on the trials when they gave the normative base-rate congruent response and thereby successfully override their default response. This supports conflict detection, and cognitive effort when overriding one’s intuitive response. Across participants and conditions there was not a general effect of larger pupil dilations following attribute presentation being linked to correct responses. Thus, the effect was specific to the incongruent condition and the group overriding their majority response. However, we note that we cannot exclude that small effects of condition and general performance may be present as we lack the power to detect them. Importantly, the presence of such effects does not exclude a separate effect of conflict detection.

Regarding pupil dilations preceding decisions, we found that base-rate responders showed larger pupil dilations in the incongruent and neutral conditions compared to the congruent condition. Indicating that pupils may dilate either when facing conflicting information, changing strategy, integrating two pieces of information, or when task difficulty is high (van der Wel & van Steenbergen, 2018). There were no differences in pupil dilations before decisions for the stereotype responders.

In the incongruent condition we did not find that pupil dilations were associated with responses across participants. However, for the stereotype responders we did find that pupil dilations in the attribute time window was associated with giving the normative response. In the decision time window, the effect was of similar size and direction, barely missing significance for the stereotype responders. Thus, it may be that pupil dilation and cognitive effort are related to giving the response opposite to one's majority response. Larger pupil dilations before decisions were associated with performance across conditions, showing a general effect of cognitive effort or attention for performance in the base-rate task.

General discussion

In this study, we used two base-rate task versions to investigate the role of attention and effort. In both versions we replicated that accuracy was lower in the incongruent than the congruent condition. The incongruent condition was also associated with longer response times compared to the congruent condition. Consistently for both versions we could classify participants into two groups based on their performance in the incongruent condition. Participants tended to mainly answer either base-rate congruently (base-rate responders) or stereotype congruently (stereotype responders). When investigating the effect of condition on response times, gaze and pupil dilation, we found significant differences due to condition in the base-rate responders. However, for stereotype responders there were no significant differences due to condition in many of these measures. Stereotype responders seem to be insensitive to, or neglect, the base-rate information on this task. These results support that "conflict detection failure" and neglecting base-rates are a significant source of "biased" responses. The base-rate responders processed both types of information as they showed slowing of responses when the information was conflicting, or the base-rates were non-informative, and their gaze changed toward the class information when the base-rates were non-informative. Thus, this group did integrate both types of information.

Response times in this study showed that the majority response given in the incongruent condition is faster compared to the minority response (stereotype congruent for stereotype responders and base-rate congruent for base-rate responders, respectively). Participants spend more time when overriding their majority response (Pennycook et al., 2015). This suggests that both stereotype and base-rate responses can be made fast and intuitively (Bago & De Neys, 2017; Pennycook et al., 2015; Raoelison et al., 2020). There were slower response times in the incongruent condition for base-rate responders in both base-rate versions, suggesting non-negligible interference from stereotype information. Previous studies have interpreted slower response times in the incongruent condition as evidence that using base-rates requires Type 2 processing. We argue against this interpretation as Pennycook et al. (2015) showed that reversing the order of the different information types (attribute/stereotype vs. base-rate) can alter which response is slower. Further, this study highlights that the majority response is a major determinant of which response is slowest. Additionally, the items selected for inclusion in a task and majority preference for a certain outcome can substantially influence and even reverse response times (Krajbich et al., 2015).

Response times in the neutral condition was slower in Base-rate gaze version but not in the pupillometry version. Base rates were presented on the first slide in the Base-rate pupillometry version compared to being displayed during the decision phase in the gaze version. Thus, participants could early understand that they were not informative and focus on the stereotype information. Slower responses in the neutral condition in the Base-rate – gaze version may be due to a change of strategy in the base-rate responders, from relying on base-rate information to using the stereotype information. Alternatively, it may be due to the novelty of the stimulus, as only a quarter of the trials had equal base-rates.

Gaze - Participants look more at class information and the option they choose

Investigating gaze times showed that participants in general tend to look more at the class information compared to the base-rate information. This may be due to the base-rates being faster to read, or the comparison between option may be fast as the relative base-rates were extreme. Participants also spend less time looking at the base-rates when they were equal. Indicating that participants quickly evaluated them as non-informative. There were group differences in relative gaze at the information types, where the base-rate responders did look more at the base-rate information. Conflicting information in the incongruent condition did not change the relative gaze time investigating the base-rate information. Possibly, the extreme base-rates made comparison of base-rates easy. Further investigation with moderate or more similar base-rates may show a different pattern. Participants tended to look substantially more at the option they ended up choosing. Potentially, gaze may have biased choices toward the attended option (Krajbich, 2019; Krajbich et al., 2015) in addition to reflect choice (Westbrook et al., 2020), or rationalization of their choice (J. St. B. T. Evans, 2019; Pennycook et al., 2015).

Drift diffusion modelling – conflicting information increases task difficulty not deliberation

For both versions the incongruent condition did not have a higher decision threshold, but rather a lower drift-rate compared to the congruent and neutral condition. Conflicting information does not increase evidence accumulation, response caution, or increase deliberation, but rather subjective task difficulty is higher in the incongruent condition. Thus, slower response times in the incongruent condition should not be interpreted as indicating Type 2 processing or deliberation. Conflicting intuitions are likely not the source of increased threshold or Type 2 processing on the base-rate task. Other factors such as change of strategy (or novelty) may increase decision threshold or engage Type 2 processing, as suggested by longer response times in the neutral condition in the Base-rate - gaze version.

The DDM yielded a response bias toward stereotype congruent responses in the Base-rate gaze version, and a response bias toward base-rate congruent responses in the Base-rate - pupillometry version (SOM). The two versions showed a reversal of which response type was fastest for the full sample. Therefore, we highlight that response bias can be altered by the task structure and may not be consistent across participants. We regard this study as evidence against the presence of a general response bias favoring stereotype responses. As task structure, choice preference, and stimulus materials can substantially alter response times we advise researchers to carefully consider these factors when investigating response biases in dual-process research. Further, we advise against relying on comparison of mean response times, and rather use tools such as evidence accumulation models which can decompose response time distributions into latent decision parameters (Myers et al., 2022).

Pupil size is related to performance, preliminary support for conflict detection and cognitive decoupling

Across conditions we found a significant difference in pupil dilation in the decision time-window. For base-rate responders we found larger pupil dilations in the incongruent and neutral conditions compared to the congruent condition. This could be interpreted as evidence supporting “cognitive decoupling” following conflicting information or change of strategy. Alternatively, it could be seen as increased effort due to higher task difficulty, as indicated by lower drift-rate, that is, the condition is more demanding due to the need to integrate two sources of information that do not converge on the same response. The neutral condition where they had to use the stereotype information may be more demanding for this group as they mainly relied on base-rate information. Indeed, updating and shifting does require cognitive effort and are linked to larger pupil dilations (Friedman & Miyake, 2017; van der Wel & van Steenbergen, 2018). However, in the incongruent condition pupil dilations before decisions were not significantly related to responses. Thus, the evidence is mixed but there is some evidence supporting the construct of cognitive decoupling, however there are valid alternative explanations.

For stereotype responders there were larger pupil dilations after attribute presentation and before decisions (missing statistical significance) when giving the base-rate congruent response. This could be seen as evidence for both “conflict detection” following the attribute presentation and “cognitive decoupling” leading to larger pupil dilations before decisions, resulting in the correct response, opposite to their majority response. The presence of pupil dilations corresponding to “conflict detection” and the following changed response, is coherent with a line of work showing that phasic LC activity (as indicated by pupil dilation) may function as a neural interrupt signal and is involved in the change of attentional set, reorienting, and cognitive flexibility (Bouret & Sara, 2005; Dayan & Yu, 2006; McGaughy et al., 2008). LC activity has been linked to proactive and preparatory processes enabling cognitive and inhibitory control, thus enabling further processing of relevant stimuli or multi-component behavior (Chmielewski et al., 2017). This separation between preparatory processes and the following decision related processes is consistent with a separation of “conflict detection” and “cognitive decoupling”.

Smaller baseline pupil size, predicted performance and indicates that participants may have had too high tonic LC activity, leading to lower task performance and higher distractibility (Aston-Jones & Cohen, 2005; Mittner et al., 2016) (SOM, analysis of pupil size and exploratory behavior). Alternatively, high neural gain may focus attention to the most salient features of the information, or in accordance with a pre-disposition (Eldar et al., 2013), leading to more focus on the stereotype information, resulting in more stereotype congruent responses (SOM).

In summary, we do find some evidence supporting the constructs of “conflict detection failure”, “conflict detection” and “cognitive decoupling” from Pennycook et al. (2015). Further, we highlight that the LC – NE system may be implicated in the implementation of these or similar mechanisms. It is possible that high tonic LC activity constrains attention and enhances salient representations, leading to errors in reasoning typically associated with Type 1 processing. Further, that intermediate LC tonic activity allows for phasic LC activity and is associated with both “conflict detection” or neural interrupt signal, and a subsequent increase in cognitive effort or “cognitive decoupling” (Type 2 processing). However, the evidence in this study can be explained without two separate processes.

That conflicting information in the incongruent condition is associated with slower response times and more cognitive effort compared to non-conflicting information in the congruent condition can be explained by higher stimulus difficulty. Slower response times in the incongruent condition may just reflect higher stimulus difficulty. Relatedly, that conflicting information is associated with larger pupil dilations as a conflict is detected is similar to many task paradigms investigating conflict detection (Laeng et al., 2011; van der Wel & van Steenbergen, 2018, but see Schacht et al., 2010). Thus, the results can be explained through other models such as sequential sampling models integrating the role of attention, gaze and the LC-NE system (Busemeyer & Townsend, 1993; Eldar, Cohen, et al., 2016; Eldar et al., 2013; Eldar, Niv, et al., 2016; Gold & Shadlen, 2007; Krajbich, 2019; Roe et al., 2001). However, the tasks in this study were not designed to test these models and it is thus considered outside the scope of this article.

Limitations

Due to errors in implementation of individual difference measures we were not able to further investigate individual differences between the two responder groups or the relationship with gaze and pupil dilation. Some data loss in pupil recording decreased statistical power, however the sample was deemed adequate to continue with analyses. There did seem to be a small learning effect in the Base-rate – Pupillometry version, however this effect was small and is not expected to have influenced the results in any substantial way (SOM section 4, Learning effects).

Conclusion

In two base-rate tasks we find evidence that participants can be separated into two groups based on their responses in the incongruent condition, stereotype responders and base-rate responders. We observed that base-rate responders were sensitive to changes across conditions in terms of response times, gaze and pupil dilation, whereas stereotype responders were not. Thus, finding evidence for "conflict detection failure" and base-rate neglect for stereotype responders. Pupil dilation results showed preliminary support for the constructs of conflict detection and cognitive decoupling. We found preliminary evidence that pupil dilations following attribute presentation may be implicated in changing responses from stereotype congruent to base-rate congruent, implicating phasic LC activity in conflict monitoring or acting as a neural interrupt signal. Additionally, we observed larger pupil dilations for base-rate responses in the incongruent condition, thus supporting the notion of cognitive effort and possibly "cognitive decoupling" being associated with base-rate congruent responses on the base-rate task. However, the results can be explained without reference to dual-process and were consistent with alternative models of decision-making.

References

Armstrong, T., & Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical Psychology Review*, *32*(8), 704–723. https://doi.org/10.1016/j.cpr.2012.09.004

Arnsten, A. F. T. (2011). Catecholamine influences on dorsolateral prefrontal cortical networks. *Biological Psychiatry*, *69*(12), e89-99. https://doi.org/10.1016/j.biopsych.2011.01.027

Aston-Jones, G., & Cohen, J. D. (2005). AN INTEGRATIVE THEORY OF LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance. *Annual Review of Neuroscience*, *28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3), 241–254. https://doi.org/10.1017/S0140525X07001653

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Bouret, S., & Sara, S. J. (2005). Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, *28*(11), 574–582. https://doi.org/10.1016/j.tins.2005.09.002

Broshek, D. K., & Barth, J. T. (2000). The Halstead-Reitan Neuropsychological Test Battery. In *Neuropsychological assessment in clinical practice: A guide to test interpretation and integration* (pp. 223–262). John Wiley & Sons, Inc.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459. https://doi.org/10.1037/0033-295X.100.3.432

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Chmielewski, W. X., Mückschel, M., Ziemssen, T., & Beste, C. (2017). The norepinephrine system affects specific neurophysiological subprocesses in the modulation of inhibitory control by working memory demands. *Human Brain Mapping*, *38*(1), 68–81. https://doi.org/10.1002/hbm.23344

Coalson, D. L., Raiford, S. E., Saklofske, D. H., & Weiss, L. G. (2010). CHAPTER 1 - WAIS-IV: Advances in the Assessment of Intelligence. In L. G. Weiss, D. H. Saklofske, D. L. Coalson, & S. E. Raiford (Eds.), *WAIS-IV Clinical Use and Interpretation* (pp. 3–23). Academic Press. https://doi.org/10.1016/B978-0-12-375035-8.10001-1

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network (Bristol, England)*, *17*(4), 335–350. https://doi.org/10.1080/09548980601004024

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. https://doi.org/10.1038/nn.3428

Eldar, E., Cohen, J. D., & Niv, Y. (2016). Amplified selectivity in cognitive processing implements the neural gain model of norepinephrine function. *Behavioral and Brain Sciences*, *39*, e206. https://doi.org/10.1017/S0140525X15001776

Eldar, E., Niv, Y., & Cohen, J. D. (2016). Do You See the Forest or the Tree? Neural Gain and Breadth Versus Focus in Perceptual Processing. *Psychological Science*, *27*(12), 1632–1643. https://doi.org/10.1177/0956797616665578

Engvik, H., & Clausen, S.-E. (2011). Norsk kortversjon av Big Five Inventory (BFI-20). *Tidsskrift for Norsk psykologforening*, *48*(9). https://psykologtidsskriftet.no/oppsummert/2011/09/norsk-kortversjon-av-big-five-inventory-bfi-20

Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Evans, J. St. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. St. B. T. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, N. J., & Wagenmakers, E.-J. (2020). Evidence Accumulation Models: Current Limitations and Future Directions. *The Quantitative Methods for Psychology*, *16*(2), 73–90. https://doi.org/10.20982/tqmp.16.2.p073

Friedman, N. P., & Miyake, A. (2017). Unity and Diversity of Executive Functions: Individual Differences as a Window on Cognitive Structure. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond “Heuristics and Biases.” *European Review of Social Psychology*, *2*(1), 83–115. https://doi.org/10.1080/14792779143000033

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In *Subjective probability* (pp. 129–161). John Wiley & Sons. https://doi.org/10.1038/scientificamerican1157-128

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*(3), 592–596. https://doi.org/10.1037/0033-295X.103.3.592

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 513–525. https://doi.org/10.1037/0096-1523.14.3.513

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

Glimcher, P. W. (2003). The Neurobiology of Visual-Saccadic Decision Making. *Annual Review of Neuroscience*, *26*(1), 133–179. https://doi.org/10.1146/annurev.neuro.26.010302.081134

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Hess, E. H., & Polt, J. M. (1964). Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Joshi, S., & Gold, J. I. (2020). Pupil size as a window on neural substrates of cognition. *Trends in Cognitive Sciences*, *24*(6), 466–480. https://doi.org/10.1016/j.tics.2020.03.005

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron*, *89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kahneman, D. (2011). *Thinking, fast and slow* (p. 499). Farrar, Straus and Giroux.

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kashdan, T. B., Stiksma, M. C., Disabato, D. J., McKnight, P. E., Bekier, J., Kaji, J., & Lazarus, R. (2018). The Five-Dimensional Curiosity Scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, *73*, 130–149. https://doi.org/10.1016/j.jrp.2017.11.011

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–17. https://doi.org/10.1017/S0140525X00041157

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*(4), 665–682. https://doi.org/10.1037/a0020198

Krajbich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, *29*, 6–11. https://doi.org/10.1016/j.copsyc.2018.10.008

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*(1), Article 1. https://doi.org/10.1038/ncomms8455

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processing*, *12*(1), 13–21. https://doi.org/10.1007/s10339-010-0370-z

Lin, H., Pennycook, G., & Rand, D. G. (2023). Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, *230*, 105312. https://doi.org/10.1016/j.cognition.2022.105312

Mækelæ, M. J., Klevjer, K., Westbrook, A., Eby, N. S., Eriksen, R., & Pfuhl, G. (2023). Is it cognitive effort you measure? Comparing three task paradigms to the Need for Cognition scale. *PLOS ONE*, *18*(8), e0290177. https://doi.org/10.1371/journal.pone.0290177

Mækelæ, M. J., & Pfuhl, G. (2019). Deliberate reasoning is not affected by language. *PLOS ONE*, *14*(1), e0211428. https://doi.org/10.1371/journal.pone.0211428

Mathot, S. (2018). Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition*, *1*(1), 16. https://doi.org/10.5334/joc.18

McGaughy, J., Ross, R. S., & Eichenbaum, H. (2008). Noradrenergic, but not cholinergic, deafferentation of prefrontal cortex impairs attentional set-shifting. *Neuroscience*, *153*(1), 63–71. https://doi.org/10.1016/j.neuroscience.2008.01.064

Mittner, M., Hawkins, G. E., Boekel, W., & Forstmann, B. U. (2016). A Neural Model of Mind Wandering. *Trends in Cognitive Sciences*, *20*(8), 570–578. https://doi.org/10.1016/j.tics.2016.06.004

Myers, C. E., Interian, A., & Moustafa, A. A. (2022). A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1039172

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154. https://doi.org/10.1037/xlm0000372

Neys, W. D., Vartanian, O., & Goel, V. (2008). Smarter Than We Think: When Our Brains Detect That We Are Biased. *Psychological Science*, *19*(5), 483–489. https://doi.org/10.1111/j.1467-9280.2008.02113.x

Pan, W., Geng, H., Zhang, L., Fengler, A., Frank, M., ZHANG, R.-Y., & Chuan-Peng, H. (2022). *A Hitchhiker's Guide to Bayesian Hierarchical Drift-Diffusion Modeling with dockerHDD*. https://doi.org/10.31234/osf.io/6uzga

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian Stochastic Modelling in Python. *Journal of Statistical Software*, *35*(4), 1–81.

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

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, *19*(3), 528–534. https://doi.org/10.3758/s13423-012-0249-3

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 544–554. https://doi.org/10.1037/a0034887

Petersen, S. E., & Posner, M. I. (2012). The Attention System of the Human Brain: 20 Years After. *Annual Review of Neuroscience*, *35*, 73–89. https://doi.org/10.1146/annurev-neuro-062111-150525

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012*.

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making*, *14*, 170–178.

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., & Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*(1), 13289. https://doi.org/10.1038/ncomms13289

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370. https://doi.org/10.1037/0033-295X.108.2.370

Salthouse, T. A. (2011). What cognitive abilities are involved in trail-making performance? *Intelligence*, *39*(4), 222–232. https://doi.org/10.1016/j.intell.2011.03.001

Schacht, A., Dimigen, O., & Sommer, W. (2010). Emotions in cognitive conflicts are not aversive but are task specific. *Cognitive, Affective & Behavioral Neuroscience*, *10*(3), 349–356. https://doi.org/10.3758/CABN.10.3.349

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Simon, H. A. (1957). *Models of man; social and rational* (pp. xiv, 287). Wiley.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168. https://doi.org/10.1016/j.tins.2004.01.006

Spencer, R. C., & Berridge, C. W. (2019). Receptor and circuit mechanisms underlying differential procognitive actions of psychostimulants. *Neuropsychopharmacology*, *44*(10), 1820–1827. https://doi.org/10.1038/s41386-019-0314-y

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Unsworth, N., & Robison, M. K. (2017). A locus coeruleus-norepinephrine account of individual differences in working memory capacity and attention control. *Psychonomic Bulletin & Review*, *24*(4), 1282–1311. https://doi.org/10.3758/s13423-016-1220-5

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The Role of Locus Coeruleus in the Regulation of Cognitive Performance. *Science*, *283*(5401), 549–554. https://doi.org/10.1126/science.283.5401.549

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

van Steenbergen, H., & Band, G. P. H. (2013). Pupil dilation in the Simon task as a marker of conflict processing. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00215

Vartanian, O., Beatty, E. L., Smith, I., Blackler, K., Lam, Q., Forbes, S., & De Neys, W. (2018). The Reflective Mind: Examining Individual Differences in Susceptibility to Base Rate Neglect with fMRI. *Journal of Cognitive Neuroscience*, *30*(7), 1011–1022. https://doi.org/10.1162/jocn_a_01264

Vehlen, A., Spenthof, I., Tönsing, D., Heinrichs, M., & Domes, G. (2021). Evaluation of an eye tracking setup for studying visual attention in face-to-face conversations. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-81987-x

Westbrook, A., van den Bosch, R., Määttä, J. I., Hofmans, L., Papadopetraki, D., Cools, R., & Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, *367*(6484), 1362–1366. https://doi.org/10.1126/science.aaz5891

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 14. https://doi.org/10.3389/fninf.2013.00014

Yang, X., & Krajbich, I. (2023). A dynamic computational model of gaze and choice in multi-attribute decisions. *Psychological Review*, *130*(1), 52–70. https://doi.org/10.1037/rev0000350

Supplementary online material (SOM)

The influence of visual attention and cognitive effort on base-rate neglect

Overview of SOM

1. Individual difference measures
2. Base-rate - gaze version
3. Base-rate - pupillometry version
4. Learning effects
5. Hierarchical drift-diffusion modelling
6. The role of the Locus Coeruleus – Noradrenaline system
7. Instructions

All reported mixed models from manuscript analyzed with the “lme4” package (Bates et al., 2015) in the statistical environment R (RStudio, version 4.1.2).

Hierarchical drift-diffusion models were implemented with the HDDM Python toolbox (HDDM version 0.9.5) using a dockerHDDM (Pan et al., 2022; Wiecki et al., 2013).

1. Individual difference measures

1.1 Materials

Thinking disposition.

Thinking disposition was measured with the abbreviated Need for Cognition scale (NFC) (Cacioppo et al., 1984). The NFC measures individual differences in the tendency to engage in and enjoy cognitively effortful activities. The scale consists of 18 items that are rated on a 5-point Likert scale from 1 = “*Extremely uncharacteristic of me*” to 5 = “*Extremely characteristic of me*”. An example item is “I find satisfaction in deliberating hard and for long hours”. The scale total ranges from 18 to 90. The scale was presented on a computer and implemented in Qualtrics (Qualtrics, Provo, UT).

Cognitive ability

Cognitive ability was measured with a battery of rational reasoning tasks, the digit-symbol substitution test and the trail-making test part-B.

Rational reasoning

A battery of rational reasoning tasks from the heuristics and bias literature was used to create a rational reasoning ability composite score. The battery included 14 items. Correct answers were scored as 1, incorrect as 0. The complete composite rational reasoning score ranged 0 to 14. From the extended Cognitive reflection test (Toplak et al., 2014) we used items 2-7. We used two fully disjunctive reasoning problems, “the marriage problem” (Levesque, 1986) and a “knight and knave problem” (Smullyan, 1978), one probability matching task (J. Koehler & James, 2010), one conditional reasoning problem (Lehman et al., 1988), one covariation detection problem (Stanovich & West, 1998), one Bayesian reasoning problem (Toplak et al., 2007; adapted from Fischoff & Beyth-Mayrom, 1983), one probability estimation problem “the bus problem” (Teigen & Keren, 2007), one medical Bayesian reasoning problem (Gigerenzer et al., 2007).

Digit Symbol Substitution Test (DSST)
The DSST is a timed (90 seconds) paper and pencil measure of processing speed. Participants have to fill in symbols that are paired to each digit (1-9) following a digit-symbol pair code. Performance is measured as the number of symbols correctly coded. The DSST may in addition to processing speed measure psychomotor speed, short-term-visual memory, attention, cognitive flexibility and motivation (Coalson et al., 2010).

Trail making test (TMT)
The TMT is a measure dependent on several mental abilities such as psychomotor speed, mental flexibility, visual scanning, and executive function (Halstead-Reitan; Tombaugh, 2004; Salthouse, 2012). The TMT consists of two parts; A and B. In part A (TMT-A) participants are instructed to draw a line between 25 dots, containing numbers from 1 -25, in ascending order. Performance is measured in seconds to complete the task (reverse scored). Part B (TMT-B) consists of 25 dots with both letters and numbers inside. Participants are instructed to draw a line in ascending order, alternating between letters and numbers (1 – A – 2 -B – 3 - C… 13) until the end. Part B scoring is the same as for Part A. We here use time on Part B (reverse scored) as a measure of cognitive ability as Part B has the highest relation to full scale intelligence and fluid intelligence (Corrigan & Hinkeldey, 1987; Salthouse, 2012).

Curiosity
Curiosity was measured with two subscales, Joyous exploration and social curiosity, from the five-dimensional curiosity scale (Kashdan et al., 2018). Joyous exploration was measured with five items and is a central part of curiosity capturing a preference for new information, new experiences and favoring self-expansion over security. Social curiosity was measured with five items and captures an interest in how other people think and behave.

Personality
Personality was measured with a short 20 item version of the big five inventory (BFI-20) (Engvik & Clausen, 2011; John & Srivastava, 1999). The inventory was included for participants (N = 30) taking part in 2 days of testing, the inventory was included on Day 1 (Mækelæ et al., 2023), but not when the base-rate tasks were completed (day 2). After the interruption (COVID-19) this scale was dropped. Given the low N, the scale was not used at all.

1.2 Procedure notes

A sample of participants (N = 30) for the study was recruited for two days of testing where day 1 was a separate project (Mækelæ et al., 2023) and day 2 was the current study. The remaining participants were recruited to take part in this study only. Due to an error in implementation of surveys in Qualtrics (Qualtrics, Provo, UT) NFC was not included for participants taking part in only 1 day of testing. Regarding the TMT, one experimenter, testing eleven participants, did not correctly instruct participants who made errors when performing the task. Therefore, a subsample of the TMT scores had to be removed. Additionally, there were exclusions due to behavioral performance and low-quality pupil recording (see main manuscript). Therefore, it was decided that the sample size for individual difference measures were too low to be reliable and was thus not included in the main manuscript. Results are reported for transparency.

Accuracy for each task is reported as number of correct responses, this includes base-rate congruent responses in the incongruent condition.

1.3 Results

Descriptive statistics for individual difference measures can be found in Table S1.

Table S1

Descriptive statistic for individual difference measures

	Mean	SD	Minimum	Maximum	N
RQ	5.27	2.27	0	10	59
NFC	59.03	12.82	24	81	40
DSST	61.93	10.39	38	86	59
TMT-B	58.30	27.13	28	149	55
Curiosity - JE	27.09	4.83	15	35	53
Curiosity - SC	24.25	6.36	7	35	53
Gaze	0.78	0.22	0.43	1	49
Pupillometry	0.74	0.18	0.44	0.99	54

Note. RQ = rational reasoning composite score, NFC = Need for cognition, DSST = Digit-symbol substitution test, TMT-b = Trail making test part B. TMT-B measured in seconds, Curiosity - JE = Curiosity - Joyous exploration subscale, Curiosity - JC = Curiosity – Social curiosity subscale. Gaze = Accuracy in base-rate – gaze version. Pupillometry = Accuracy in base-rate – pupillometry version.

Correlations between individual difference measures and accuracy in both base-rate tasks are presented in Table S2.

Table S2

Base-rate tasks and correlations with individual difference measures

		NFC	RQ	DSST	TMT-B	Curiosity - JE	Curiosity - SC
Gaze		N =36	N = 48	N = 48	N = 47	N = 46	**N = 47**
	Corr.	.03	.27	.23	.23	-.09	**.32**
	p-value	.885	.066	.124	.116	.537	**.028**
Pupil		N = 38	N = 54	N = 54	N = 51	N = 52	**N = 52**
	Corr.	.25	.15	.06	-.22	-.06	**.33**
	p-value	.127	.268	.687	.125	.663	**.017**
NFC			N = 36	N = 36	N = 36	**N = 35**	N = 36
	Corr.		.18	.21	-.03	**.67**	-.05
	p-value		.304	.222	.886	**<.001**	.776
RQ				**N = 48**	N = 47	N = 46	N = 47

	Corr.	**.42**	-.13	.09	.01
	p-value	**.003**	.369	.549	.961
DSST			**N = 47**	N = 46	N = 47
	Corr.		**-.35**	.02	-.00
	p-value		**.016**	.977	.977
TMT-B				N = 45	N = 46
	Corr.			.05	-.13
	p-value			.740	.402
C - JE					N = 46
	Corr.				-.05
	p-value				.742

Legend. Corr. = Spearman's rho. Significant correlations in bold, p < .05. Gaze = Accuracy in the Base-rate -gaze version. Pupil = Accuracy in the Base-rate -pupillometry version. NFC = Need for cognition scale, RQ = rational reasoning composite score, DSST = Digit-symbol substitution test, TMT-B = Trail making test part B, Curiosity - JE (C-JE) = Curiosity - Joyous exploration subscale, Curiosity - JC = Curiosity – Social curiosity subscale.

2. Base-rate - gaze version

Gaze at option predicts choice

We investigated whether proportional gaze time at the correct vs. the incorrect option (both group and base-rate information) was different across the three conditions. A Kruskal-Wallis one-way ANOVA with Dunn's post-hoc comparisons showed that there was a significant group difference across conditions (X^2 = 170.650, p < .001, df = 2), and that there was a significant difference between all groups (all p-values < 0.01). Participants spent the most time looking at the correct option in the congruent condition, followed by the neutral condition, and the incongruent condition had the lowest proportional gaze time at the correct option (Fig. 6). However, the response groups did differ with "base-rate responders" looking most at the correct option in the congruent, then incongruent and lastly neutral condition (X^2 = 114.589, p < .001, df = 2, Dunn's post-hoc test all pair-wise comparisons p < .001). "Stereotype responders" looked most at the correct option in the congruent, then neutral and lastly incongruent condition (X^2 = 246.765, p < .001, df = 2, Dunn's post-hoc comparisons all but congruent vs neutral pair-wise comparisons p < .001, Fig. 3 right-hand panel in main manuscript).

Next, we in looked at proportional gaze in the incongruent condition, comparing correct (base-rate congruent) responses and incorrect (stereotype congruent) responses. Correct responses were associated with significantly more time looking at the correct option compared to incorrect responses (U = 115460.500, d = -0.66, p < .001). Participants giving the correct response (M = 0.70, SD = 0.23) proportionally looked more at the correct option, whereas participants giving the incorrect response (M = 0.33, SD = 0.31) tended to look more at the incorrect option (both significantly different from 0.5 one-sided t-test p's < .001).[1]

Gaze at option type and information type predicts choice in incongruent condition

To assess the relative influence of gaze at information type (class vs. base-rate, information split on screen between top and bottom) and choice options (correct vs. incorrect option, information split on screen left and right) on responses in the incongruent condition, we conducted generalized linear mixed models (GLMM's). A model with response (correct vs. incorrect) as outcome and proportional gaze at information type (class > 0.5 > base-rate) as a fixed factor and participants as a random factor showed that gaze at information type was a significant predictor of response (b = -1.10, SE = 0.35, Z = -3.11, p = 0.002) where gaze at information type accounted for 0.6% of the variance in responses (Table S3). A model with response as outcome and proportional gaze at correct option as a fixed factor and participants as random factor showed that proportional gaze at correct option was a significant predictor of responses (b = 5.16, SE = 0.44, Z = 11.84, p < .001) and gaze at option accounted for 12.4% of the variance in responses (Table S4). When including both fixed factors as predictors of responses proportional gaze at group information was no longer a significant predictor (p = 0.094), whereas gaze at correct option still was (p < .001), see Table S5. These results indicate that gaze at information type is associated with responses, but the effect is small. Gaze at options on the other hand has a medium to large association with choice.

Table S3

Generalized linear mixed model with response as outcome, proportional gaze at information type as fixed factor and participants as random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	2.60	1.18; 4.02	3.59	<.001
Proportional gaze – Information Type	-1.10	-1.79; -0.41	-3.11	.002
Random Effects				
σ^2	3.29			
$\tau_{00\ subj\_idx}$	16.45			
ICC	0.83			
N $_{subj\_idx}$	49			
Observations	1696			
Marginal R^2 / Conditional R^2	0.006 / 0.834			

Table S4
Generalized linear mixed model with response as outcome, proportional gaze at choice option as fixed factor and participants as random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	-0.97	-2.33; 0.40	-1.39	.165
Proportional gaze – Choice option	5.16	4.31; 6.01	11.84	<.001
Random Effects				
σ^2	3.29			
$\tau_{00 \text{ subj\_idx}}$	15.32			
ICC	0.82			
N $_{\text{subj\_idx}}$	49			
Observations	1696			
Marginal R^2 / Conditional R^2	0.124 / 0.845			

Table S5
Generalized linear mixed model with response as outcome, proportional gaze at information type and option type as fixed factors and participants as random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	-0.47	-1.95; 1.02	-0.62	.538
Proportional gaze – Information Type	-0.72	-1.55; 0.12	-1.68	.094
Proportional gaze – Choice option	5.11	4.25; 5.97	11.70	<.001
Random Effects				
σ^2	3.29			
$\tau_{00 \text{ subj\_idx}}$	15.30			
ICC	0.82			
N $_{\text{subj\_idx}}$	49			
Observations	1696			
Marginal R^2 / Conditional R^2	0.128 / 0.846			

Both Stereotype responders and Base-rate responders look more at the option they choose.”

Stereotype responders looked more at the correct (base-rate congruent) option in the incongruent condition when giving the correct response, *M* = 0.70, *SD* = 0.22 and more at the incorrect option when giving the incorrect response, *M* = 0.32, *SD* =0.31, and this difference was significant U = 7666.000, d = -0.67, p < .001.

Base-rate responders looked more at the correct option when giving the correct response in the incongruent condition, *M* = 0.70, *SD* = 0.23 and more at the incorrect option (stereotype congruent) when giving the incorrect response, *M* = 0.41, *SD* = 0.27, and this difference was significant U = 19443.500, d = -0.60, p < .001).

3. Base-rate - pupillometry version

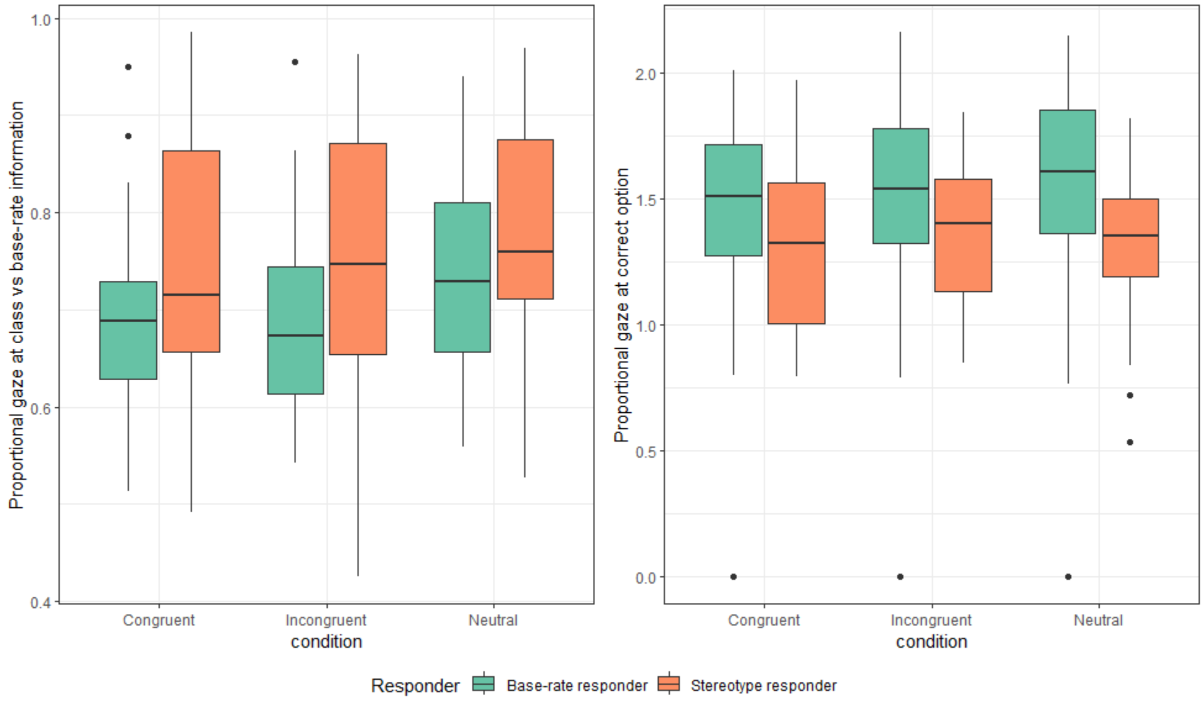
In the Base-rate - pupillometry version the base-rate responders looks more at the base-rate information, similar to results from Base-rate – gaze version.

We assessed if there were differences in gaze at information type (class vs. base-rate) across conditions and groups (stereotype- and base-rate – responders), as we found in base-rate – gaze version. We note that the task structure differs between the two tasks. In the Base-rate - pupillometry version, gaze is measured before the attribute information, whereas gaze is measured after attribute information in the Base-rate – gaze version.

A Kruskal-Wallis one-way ANOVA showed there was a significant difference in gaze at information type for the three conditions (χ^2 = 22.123, p < .001, df = 2). Dunn’s post hoc comparisons showed that participants spend less time looking at the non-informative base-rates in the neutral condition compared to the congruent (p < .001) and incongruent (p < .001) conditions. Similar to the Base-rate - gaze version there were no differences in gaze at information type in the congruent and incongruent conditions (p = .860).

Figure S1

Proportional gaze at information type (class vs base rate) and proportional gaze at correct option separately for Stereotype responders and Base-rate responders.



Comparing stereotype responders and base-rate responders we find that base-rate responders (*M* = 0.70, *SD* = 0.22) look significantly more at the base-rate information compared to stereotype responders (*M* = 0.75, *SD* = 0.25, U = 1.275, d = 0.06, p = .005). When analyzing stereotype responders and base-rate responders separately we find no difference in the three conditions for the stereotype responders (χ^2 = 1.280, p = .527, df = 2). However, base-rate responders showed a significant group difference (χ^2 = 26.741, p < .001 .527, df = 2). Dunn's post hoc comparisons showed that again the neutral condition is significantly different from the congruent (p < .001) and incongruent (p < .001) conditions, but there is no difference between the congruent and incongruent conditions (p = .410). Thus, replicating that base-rate responders look more at the base rate information and seem to be sensitive to changes due to condition whereas stereotype responders are not.

Generalized mixed models reported in Base-rate - pupillometry version

From main text: "To test if pupil size is predictive of performance on the Base-rate pupillometry task across conditions we conducted a series of GLMM's with response as outcome, condition and pupil measures (trial-baseline, attribute and decision) as fixed factors and participants as random factor."

Table S6

Generalized linear mixed model with response as outcome, condition and trial baseline pupil size as fixed factors and participants as random factor

		Response		
Predictors	*Estimates*	*CI*	*Statistic*	*p*

(Intercept)	3.46	2.91; 4.01	12.29	<.001
Condition – Incongruent	-2.75	-3.16; -2.35	-13.35	<.001
Condition – Neutral	-0.88	-1.33; -0.42	-3.79	<.001
Trial Baseline	-0.25	-0.37; -0.14	-4.33	<.001
Random Effects				
σ^2	3.29			
$\tau_{00\ subj\_idx}$	1.51			
ICC	0.32			
N $_{subj\_idx}$	38			
Observations	2444			
Marginal R^2 / Conditional R^2	0.235 / 0.476			

Table S7

Generalized linear mixed model with response as outcome, condition and pupil dilation in attribute time window as fixed factors and participants as random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	3.49	2.92; 4.07	11.95	<.001
Condition – Incongruent	-2.80	-3.23; -2.36	-12.61	<.001
Condition – Neutral	-0.94	-1.43; -0.45	-3.79	<.001
Pupil dilation – Attribute	0.10	-0.02; 0.22	1.56	.119
Random Effects				
σ^2	3.29			
$\tau_{00\ subj\_idx}$	1.50			
ICC	0.31			
N $_{subj\_idx}$	38			
Observations	2142			
Marginal R^2 / Conditional R^2	0.232 / 0.473			

Table S8

Generalized linear mixed model with response as outcome, condition and pupil dilation in decision time window as fixed factors and participants as random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	3.42	2.86; 3.98	11.92	<.001
Condition – Incongruent	-2.69	-3.12; -2.26	-12.36	<.001
Condition – Neutral	-0.92	-1.39; -0.44	-3.77	<.001
Pupil dilation – Decision	0.15	0.03; 0.27	2.37	.018
Random Effects				
σ^2	3.29			
$\tau_{00\ subj\_idx}$	1.45			
ICC	0.31			
N_{subj_idx}	38			
Observations	2145			
Marginal R^2 / Conditional R^2	0.221 / 0.459			

Table S9

Generalized linear mixed model with response as outcome, condition, trial baseline pupil size and pupil dilation in attribute time window and their interaction as fixed factors and participants as random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	3.53	2.95; 4.11	11.92	<.001
Condition – Incongruent	-2.81	-3.25; -2.37	-12.60	<.001
Condition – Neutral	-0.92	-1.41; -0.43	-3.68	<.001
Trial Baseline	-0.29	-0.45; -0.13	-3.61	<.001
Pupil dilation – Attribute	-0.09	-0.24; 0.07	-1.12	.264
Baseline:Attribute	0.03	-0.06; 0.12	0.64	.523
Random Effects				

σ^2	3.29
$\tau_{00\ \text{subj\_idx}}$	1.53
ICC	0.32
N $_{\text{subj\_idx}}$	38
Observations	2142
Marginal R^2 / Conditional R^2	0.241 / 0.482

Table S10
Generalized linear mixed model with response as outcome, condition, trial baseline pupil size and pupil dilation in decision time window and their interaction as fixed factors and participants as random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	3.47	2.90; 4.04	11.89	<.001
Condition – Incongruent	-2.71	-3.14; -2.28	-12.34	<.001
Condition – Neutral	-0.91	-1.39; -0.43	-3.71	<.001
Trial Baseline	-0.29	-0.44; -0.14	-3.80	<.001
Pupil dilation – Decision	-0.03	-0.18; 0.12	-0.41	.685
Baseline:Decision	0.04	-0.06; 0.13	0.76	.448
Random Effects				
σ^2	3.29			
$\tau_{00\ \text{subj\_idx}}$	1.49			
ICC	0.31			
N $_{\text{subj\_idx}}$	38			
Observations	2145			
Marginal R^2 / Conditional R^2	0.231 / 0.471			

A model with a three-way interaction between pupil measures in the trial baseline-, attribute- and decision- time windows did not converge.

4. Learning effects

Base-rate – gaze version

To investigate if there were any learning effects in the Base-rate – gaze version we compared trial number by correct and incorrect responses in the incongruent condition. A logistic regression showed that there was no significant difference in trial number between correct and incorrect responses χ^2 (1767) = 0.372, p = .372. Thus, we found no evidence of learning effects in the Base-rate – gaze version.

Base-rate – pupillometry version

To investigate if there were any learning effects in the Base-rate – pupillometry version we compared trial number by correct and incorrect responses in the incongruent condition. A logistic regression showed that there was a significant difference in trial number between correct and incorrect responses χ^2 (2021) = 20.789, p < .001. The model explained 1% of the variance in responses, and trial number was a significant predictor of correct responses (β = 0.009, SE = 0.002, z = 4.53, p < .001. The results show that there was a small learning effect in the Base-rate – pupillometry version where participants responded more correct in the incongruent condition as the task progressed. The effect was very small, 1% of variance in responses, thus no further steps were considered necessary to correct for learning effects.

5. Hierarchical drift-diffusion modelling

We used hierarchical drift-diffusion modelling to investigate if conflicting responses engages Type 2 processing. In the drift-diffusion model an increased threshold would indicate increased information sampling and more cautious responding and thus corresponds to Type 2 processing (Lin et al., 2023).

For each model we ran 5 Markov chains with 20,000 samples each, 12,000 of which were burn-in. Every second sample was discarded as thinning to reduce autocorrelation in chains. For the models looking at bias in the incongruent condition, sample size was increased to 35,000 with 20,000 as burn-in and every second sample discarded. Model convergence was assessed with visual inspection of the trace, autocorrelation, the marginal posterior, and the Gelman-Rubin R statistic. All parameters had an R-hat value below 1.05. Model comparison was conducted with the deviance information criterion (DIC). Lower DIC indicates better fit. See model comparison for Base-rate – gaze version in Table S11 and model comparison for Base-rate – pupillometry version in Table S12.

Table S11
Hierarchical Drift-Diffusion Model Comparison Base-Rate – gaze version

M11	Null model	DIC	7166
M12	Separate threshold	DIC	6910
M13	Separate drift-rate	DIC	5748
M14	Separate drift-rate and threshold	DIC	5755

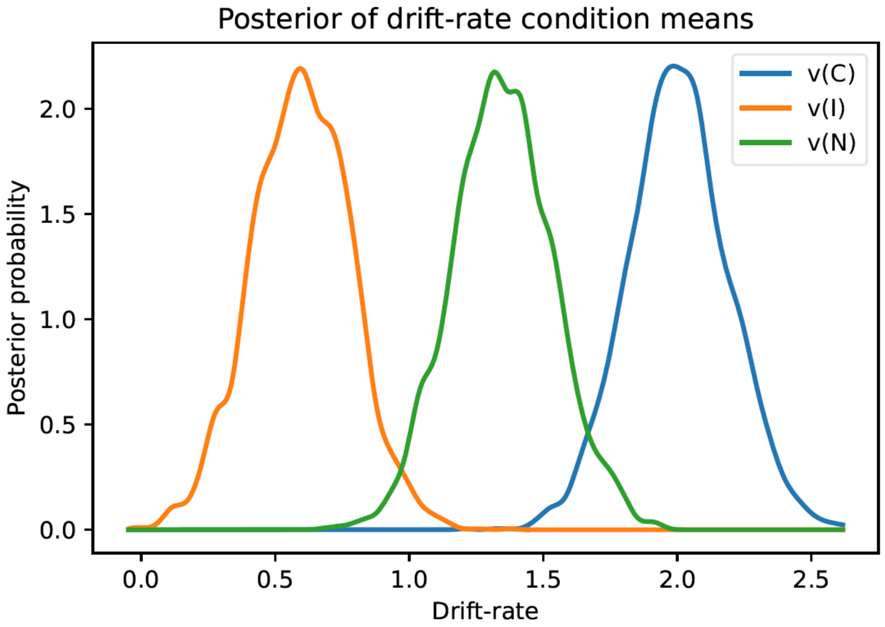
Table S12
Hierarchical Drift-Diffusion Model Comparison Base-Rate – pupillometry version

M21	Null model	DIC	9728
M22	Separate threshold	DIC	9826
M23	Separate drift-rate	DIC	8469
M24	Separate drift-rate and threshold	DIC	8510

For both tasks the model with the best fit for the data was a model with separate drift-rate for each condition (M13 and M23, Fig S2 and S4). Additionally, as seen from Figure S3 and Figure S5 (models with both separate drift-rate and threshold by condition, second best fit) the threshold in the incongruent condition was not higher compared to the other conditions. Thus, we find no evidence supporting higher decision threshold due to conflicting information in the two base-rate tasks.

For transparency we present visual representation of the posterior distribution for drift-rates for models M13 and M23 (separate drift-rate by condition, best fit), and the posterior distribution for thresholds for models M14 and M24 (separate threshold and drift-rate by condition, second best fit).

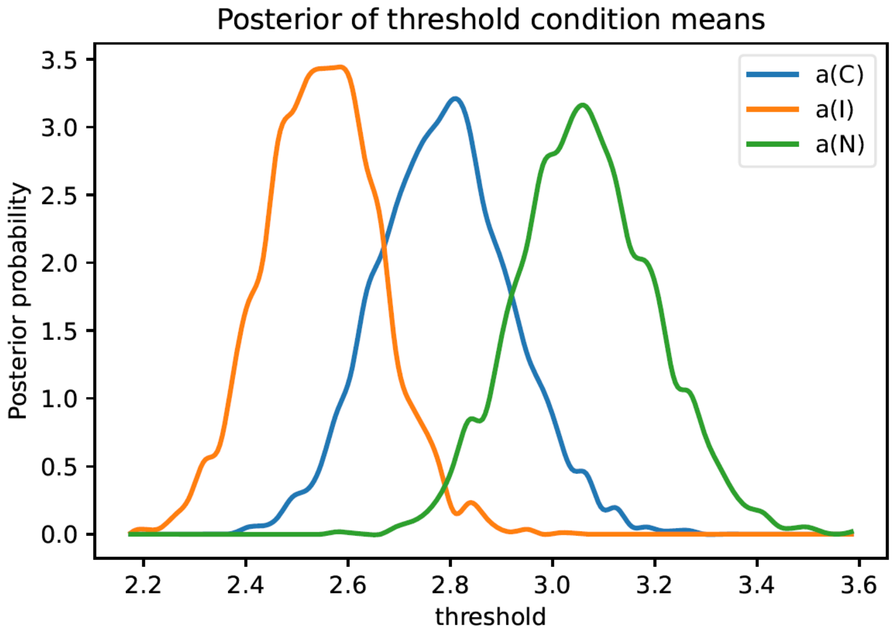
Figure S2
Posterior distribution of drift-rates in M13 – Base-rate – gaze version, separate drift-rate by condition.



Legend. V = Drift-rate. V(C) = Drift rate in congruent condition. v(I) = Drift rate in incongruent condition. v(N) = Drift rate in neutral condition.

Figure S3

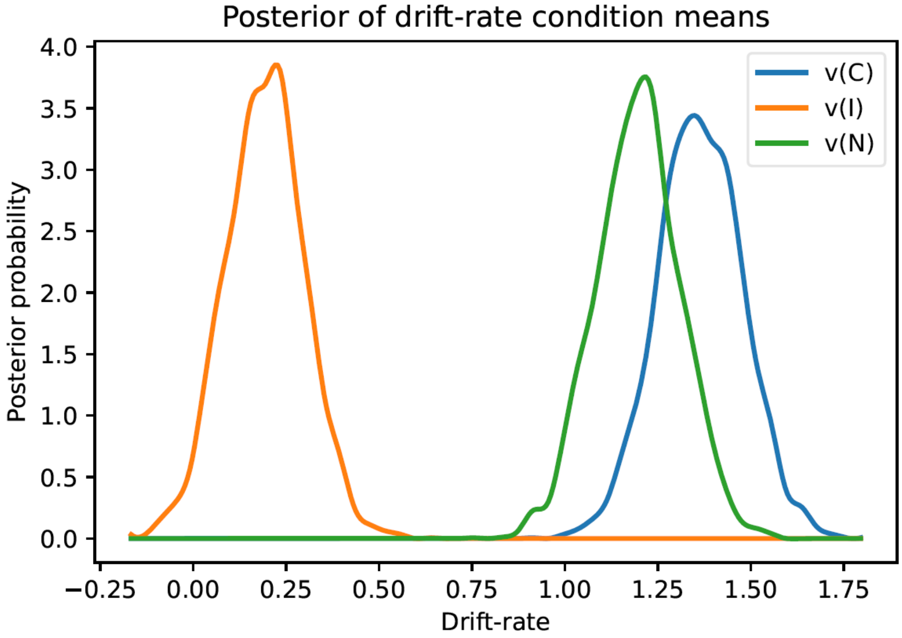
Posterior distribution of threshold in M–4 - Base-rate – gaze version, separate drift-rate and threshold by condition



Legend. a = Threshold, a(C) = Threshold in congruent condition. a(I) = Threshold in incongruent condition. a(N) = Threshold in neutral condition.

Figure S4

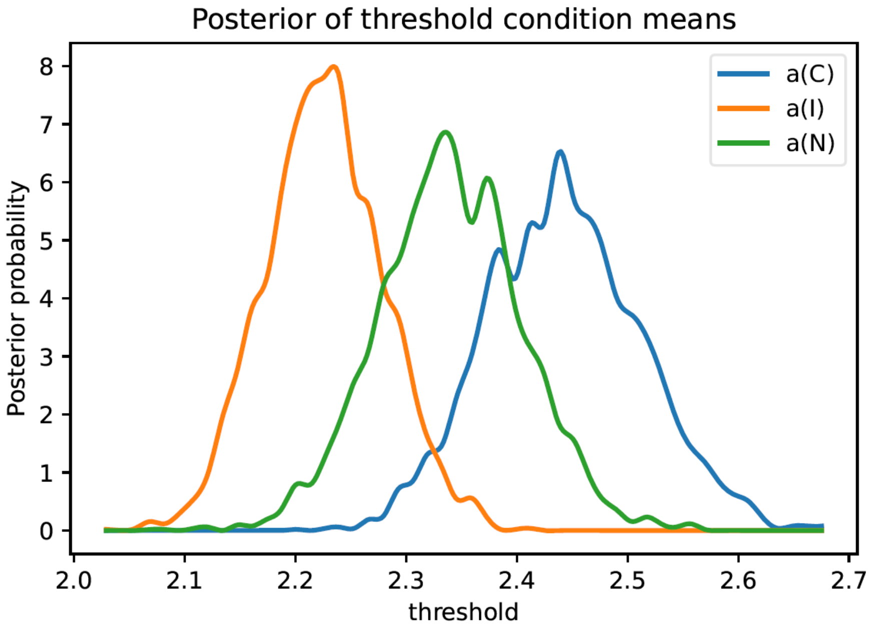
Posterior distribution of drift-rates i– M23 - Base-rate – pupillometry version, separate drift-rate by condition.



Legend. v = Drift-RI. v(C) = Drift rate in congruent condition. v(I) = Drift rate in incongruent condition. v(N) = Drift rate in neutral condition.

Figure S5

Posterior distribution of thresholds in M24 Base-rate – pupillometry version, separate drift-rate and threshold by condition.



Legend. a = Threshold. a(C) = Threshold in congruent condition. a(I) = Threshold in incongruent condition. a(N) = Threshold in neutral condition.

Bias is dependent on task structure

To assess if there is a response bias towards stereotype congruent responses we applied a stimulus coded DDM for responses in the incongruent condition. Response boundaries indicated stereotype congruent (0) and base-rate congruent (1) responses. A bias parameter of 0.5 indicates no bias. Bias parameter values < 0.5 indicate a bias towards stereotype congruent responses. Bias parameter values > 0.5 indicate a bias towards base-rate congruent responses.

Base-rate – gaze version

For the Base-rate – gaze version visual inspection of trace and autocorrelation showed minor issues. Gelman-Rubin statistic was < 1.034. Thus, there may be minor convergence issues and interpretation of results should thus be made with caution. A model with a bias parameter included (DIC = 2964) provided slightly better fit compared to a model without bias for the incongruent condition (DIC = 2971). The model had an estimated bias parameter of 0.43 with the full posterior distribution below 0.5, indicating a response bias toward the stereotype congruent response. As all information needed to produce a stereotype (class and attribute information) had been presented before the response slide, there is good reason that a response bias toward the stereotype congruent answer was present in the base-rate – gaze version.

Base-rate – pupillometry version

For the Base-rate – pupillometry version a model with a bias parameter (DIC = 4635) provided a slightly improved fit compared to a model without a bias parameter (DIC = 4638). The bias parameter was estimated at 0.52 (SD = 0.04) indicating a small bias towards base-rate congruent responses. The posterior distribution of the bias parameter showed a small overlap with 0.5, indicating that the bias may not be significantly different from 0.5. Thus, we find some evidence for a response bias for base-rate congruent responses. As the base-rate information was presented before the attribute information, evidence accumulation towards the base-rate congruent response early in the trial may have biased the responses in this task toward the base-rate congruent response.

Modelling responses in the incongruent condition with a stimulus coded DDM yielded a response bias toward stereotype congruent responses in the Base-rate gaze version, and a response bias toward base-rate congruent responses in the Base-rate - pupillometry version. Therefore, we highlight that response bias can be altered by the task structure and may not be consistent across participants. We therefore regard this study as evidence against the presence of a general response bias favoring stereotype responses. As task structure, individual differences, and selection of stimulus materials can substantially alter response times we advise researchers to carefully consider these factors when investigating response biases in dual-process research. Further, we advise against relying on comparison of mean response times, and rather use tools such as evidence accumulation models which can decompose response time distributions into latent decision parameters.

6. The role of the Locus Coeruleus – Noradrenaline system

In the current study pupil dilation was primarily included as a measure of cognitive effort (Hess & Polt, 1964; Kahneman & Beatty, 1966; van der Wel & van Steenbergen, 2018), however pupil dilation also reflects changes in the Locus Coeruleus (LC) – Norepinephrine (NE) system (Aston-Jones & Cohen, 2005; Eldar et al., 2016; Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; Joshi et al., 2016; Reimer et al., 2016). The LC-NE system is a candidate mechanism involved in “conflict detection” and possibly “cognitive decoupling”. The LC has widespread connections to most of the forebrain and modulates neural gain (Aston-Jones & Cohen, 2005; Chandler et al., 2016; Eldar et al., 2013; Waterhouse & Chandler, 2014). The LC is functionally connected to areas involved in both conflict monitoring and cognitive control, as well as working memory and executive functions, hereunder, the dorsomedial prefrontal cortex, including the anterior cingulate cortex, and the lateral prefrontal cortex (Arnsten, 2011; Botvinick et al., 2001: Botvinick et al., 2004; Joshi et al., 2016; Joshi & Gold, 2002; Shenhav et al., 2013; Spencer & Berridge, 2019; Unsworth & Robison, 2017; Usher et al., 1999). The LC-NE system has previously been linked to cognitive flexibility, attentional set shifting, inhibitory control processes, reorienting, surprise, working memory, and sustained attention (Bouret & Sara 2005; Corbetta et al., 2008; Dayan & Yu, 2006; McGaughy et al., 2008; Poe et al., 2020; Preuschoff et al., 2011; Sara & Bouret, 2012; Spencer & Berridge, 2019; Wolff et al., 2018). According to the adaptive gain theory (Aston-Jones & Cohen, 2005), the LC-NE system regulates behavioral change in the adaptive dilemma between exploiting and exploring the environment. The adaptive gain theory proposes that high tonic

LC activity (indicated by larger baseline pupil size) is related to exploratory behavior, decreased on task performance, and is marked by a reduction in phasic activity. Intermediate levels of tonic LC activity (indicated by smaller baseline pupil size) on the other hand is related to be higher on task performance and the presence of phasic activity (larger pupil dilations). Notably, phasic LC activity has also been proposed as a neural interrupt or a network reset signal (Bouret & Sara, 2005; Dayan & Yu, 2006), thus a possible mechanism for "conflict detection". Further, the LC-NE system may modulate breadth of attention such that high gain focuses attention on information that is salient, or one is predisposed to attend (in this study, salient stereotypes), whereas low gain broadens attention and increases information sampling (Eldar et al., 2013).

In the main manuscript we found that smaller baseline pupil size and larger pupil dilations before decisions were predictive of normative responding, whereas pupil dilations following attribute presentation was not. These results are consistent with the adaptive gain theory prediction that intermediate baseline pupil size is linked to higher on-task performance and that high tonic LC activity is related to decreased task performance. However, an alternative explanation for larger baseline being linked to decreased performance is that high neural gain may have enhanced the salient stereotype information and constrained attention, leading to stereotype congruent responses in the incongruent condition. Note that this applies to the incongruent condition in particular and should not influence the congruent or neutral conditions.

To test the two hypotheses: 1. Exploratory behavior and distractibility leads decreased on-task performance as predicted by the adaptive gain theory. 2. Constrained attention and enhanced salience of the stereotype information leads to stereotype responding in the incongruent condition. We assessed independently for the three conditions whether performance and trial-baseline pupil size is associated. The adaptive gain theory predicts that the relationship should be present for all three conditions (note, task difficulty is a possible confounder). Constrained attention predicts that the relationship between high baseline pupil size and performance should only be present in the incongruent condition.

We conducted three separate GLMM's (one for each condition), predicting response with trial-baseline pupil size as a fixed factor and participants as a random factor, see results in Table S11 – S13.

Table S11
Incongruent condition – GLMM predicting response with trial-baseline pupil size as a fixed factor and participants as a random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	0.74	0.05; 1.42	2.11	.035
Trial Baseline	-0.42	-0.58; -0.26	-5.13	<.001
Random Effects				

σ^2	3.29			
$\tau_{00 \text{ subj\_idx}}$	4.31			
ICC	0.57			
N $_{\text{subj\_idx}}$	38			
Observations	1260			
Marginal R^2 / Conditional R^2	0.022 / 0.576			

Table S12
Congruent condition – GLMM predicting response with trial-baseline pupil size as a fixed factor and participants as a random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	2.93	2.43; 3.42	11.57	<.001
Trial Baseline	0.02	-0.35; 0.39	0.09	.928
Random Effects				
σ^2	3.29			
$\tau_{00 \text{ subj\_idx}}$	0.31			
ICC	0.08			
N $_{\text{subj\_idx}}$	38			
Observations	576			
Marginal R^2 / Conditional R^2	0.000 / 0.085			

Table S13
Neutral condition – GLMM predicting response with trial-baseline pupil size as a fixed factor and participants as a random factor

	Response			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	2.67	2.04; 3.29	8.32	<.001
Trial Baseline	-0.03	-0.29; 0.23	-0.20	.843
Random Effects				
σ^2	3.29			
$\tau_{00 \text{ subj\_idx}}$	1.82			
ICC	0.36			
N $_{\text{subj\_idx}}$	38			
Observations	608			

Marginal R^2 / Conditional R^2	0.000/ 0.357

The results showed that trial baseline pupil size was a significant predictor of performance in the incongruent condition, but not in the congruent or neutral conditions. The results are thus consistent with the hypothesis that high neural gain constrains attention and enhances the salient stereotype information, resulting in more stereotype congruent responses in the incongruent condition. Thus, base-rate neglect may in part be influenced by high neural gain that focuses attention on the salient stereotype information, the representation is further enhanced by high neural gain at the cost of integrating more information, leading to a failure to integrate the base-rate information. This mechanism of reasoning errors on the base-rate task is similar to the Type 1 processing errors proposed by the default-interventionist account. Here Type 1 processing favors stereotype information, processes this information fast, ignores the base-rate information, and gives the stereotype congruent response.

We note that the results from these supplementary analyses are not inconsistent with the adaptive gain theory, as the two hypotheses are consistent with the same framework. Rather, the analyses aimed to disentangle the relationship between pupil size and performance in this particular task. Further, we cannot exclude that that high baseline pupil size was related to decreased performance, but higher decision difficulty led to worse performance in the incongruent condition, but not in the easier conditions.

An additional prediction from the adaptive gain theory is that higher tonic LC activity is related to more exploratory behavior. To assess this, we tested whether higher task baseline pupil size, as a measure of high tonic LC activity, was related to more gaze changes between areas of interests (AOI's), as a measure of exploratory behavior.

For this analysis we relied on data collected for the Base-rate – pupillometry version. Task baseline pupil size measured before the task had a mean of 35.53 and a standard deviation of 3.64. Gaze change was measured as the total number of recorded gaze changes between AOI's during the presentation of both class and base-rate information in the beginning of each trial in the Base-rate – pupillometry version task. Mean gaze change was calculated separately for each participant, with a group mean change between AOI's of 7.69 (*SD* = 2.20). A Pearson's correlation showed that task baseline pupil size had a positive medium correlation with gaze change (*r* = .35, p = .012). The results show that larger baseline pupil size was indeed related to more exploratory behavior measured as gaze changes. This supports that high tonic LC activity is related to exploratory behavior as predicted by the adaptive gain theory of LC function.

Instructions

"In a big research project a large number of studies were carried out where short personality descriptions of the participants were made. In every study there were participants from two population groups (e.g., politicians *and nannies).*

In each study one participant was drawn at random from the sample. You'll get to see a personality trait for this randomly chosen participant. You'll also get information about the composition of the population groups tested in the study in question.

You'll be asked to indicate to which population group the participant most likely belongs. Please answer the problems as quickly and accurately as possible."