DOKTORI (PhD) DISSZERTÁCIÓ

Szászi Barnabás Imre

Investigating human decision-making: from processes to applications

2018

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
PEDAGÓGIAI ÉS PSZICHOLÓGIAI KAR

Szászi Barnabás Imre

Investigating human decision-making: from processes to applications

Doctoral School of Psychology,
Head of the School: **Dr. Zsolt Demetrovics**, professor, Eötvös Loránd University

Behavioral Psychology Program
Head of the Programme: **Dr. Anna Veres-Székely,** professor, Eötvös Loránd University

Supervisor
Dr. Balázs Aczél, associate professor, Eötvös Loránd University

Committee members:

Dr. Klára Faragó, professor emerita, Eötvös Loránd University, President
Dr. Emese Józsa, assistant professor, Eötvös Loránd University, Secretary
Dr. Richárd Szántó, associate professor, Corvinus University of Budapest, Opponent
Dr. László Mérő, professor, Eötvös Loránd University of Budapest, Opponent
Dr. Péter Simor, assistant professor, Eötvös Loránd Tudományegyetem, Member
Dr. Dezső Németh, professor, Eötvös Loránd Tudományegyetem, Member
Dr. Péter Pajkossy, assistant professor, Budapesti Műszaki Egyetem, Member

Budapest, 2018

Table of Contents

Acknowledgements

First and foremost, I would like to express my gratitude to my mentor and friend, Balazs Aczel. Without his constant guidance, friendly pressure and benign nudging, I couldn't and wouldn't have been able to accomplish this work.

I have been also receiving a lot of mental, physical and metaphysical support from my lab mates and co-authors. Here, I would like to highlight the special contribution of Bence Palfi, Aba Szollosi, Bence Bago, Anna Palinkas, Pascal Kieslich, Péter Szécsi, Marton Kovacs and Zoltan Kekecs.

Last, but not least, I am grateful to my wife, Melinda Szaszi-Szrenka and to other members of my family. Melinda's continuous support and patience was the life-giving water for me in the desert of scientific knowledge.

List of publications that the dissertation is based upon

- Szaszi, B., Szollosi, A., Palfi, B., Aczél, B. (2017). The Cognitive Reflection Test Revisited: exploring the ways individuals solve the test. *Thinking & Reasoning*, *23*(3), 1–28. (Study 1)
- Szaszi, B., Palfi, B., Szollosi, A., Kieslich, P. Aczél, B. (2018). Thinking dynamics and individual differences: mouse-tracking analysis of the denominator task. *Judgment and Decision-making 13*(1), 23. (Study 2)
- Aczél, B., Szollosi, A., Pálfi, B., Szaszi, B., Kieslich, P. (in-press). Is Action Execution Part of the Decision-making Process? An investigation of the Embodied Choice Hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Study 3)
- Szaszi, B., Palinkas, A., Szollosi, A., Palfi, B., Aczél, B. (2017). A Systematic Scoping Review of Choice Architecture Interventions: towards a better understanding why and when nudges work. *Journal of Behavioral Decision* Making. Advance online publication. doi:10.1002/bdm.2035 (Study 4)

Each co-author has granted permission for the publications to be included in the current dissertation.

Foreword to the introduction

I'm feeling both fascinated and uncertain about the discoveries of decision science. Both feelings guided my curiosity and scientific interest.

Feeling fascinated. Decision researchers, including two recent Nobel laurates[1], have accumulated a considerable amount of knowledge building on which they could predict with great confidence in many situations how people think and decide. Consider for a moment the following question: If a bat and a ball cost $1.10 together, and the bat costs $1.00 more than the ball, how much does the ball cost? Based on previous findings, researchers could say that it is very probable that the (incorrect) 10 cent answer popped into your mind when you read the question (Frederick, 2005). Experts could also suggest ways to change the organ donation systems in several countries potentially resulting in a tenfold increase in the number of organ donations (Johnson & Goldstein, 2003). Decision researchers worked together with governments, so-called behavioral insight units where launched to apply the insights of psychology and decision-making research to improve people's life. These units reportedly achieved huge successes: increased tax revenues, school attendance rate or the number of organ donations (e.g., Behavioural Insights Ltd, 2015).

Feeling uncertain: For another moment, consider the large-scale replication project in which only thirty-six percent of replications had significant results (Open Science Collaboration, 2015), widely cited theories such as social priming or unconscious thought being questioned (Mathur et al, 2017; Bouwmeester et al., 2017), simulations suggesting that a not ignorable portion of classical findings are false positive (Ioannidis, 2005) or the behavioral interventions which happened to work in one context and country but failed to have an effect in another context or country (e.g., Jachimowicz, Duncan, Weber & Johnson, 2017).

This inseparable mix of successes and failures turned my interest towards the foundations of the field and made me examine the basic assumptions of some key theories and the applied scientific practices.

[1]Daniel Kahneman and Richard Thaler received the Nobel prize in 2002 and 2017, respectively.

Introduction

The investigation of the factors influencing human decision-making and the exploration of their underlying cognitive processes form the basis of this dissertation. In the first three studies, I examine how two factors, the first quick thoughts coming to mind and the action dynamics of thinking, influence individuals' decision-making. By tracking and manipulating the underlying cognitive processes, I tested the predictions of the default-interventionist and the embodied choice theories. One goal of decision-making research is to apply the results of the fundamental research in applied setting. Accordingly, extending the scope of the dissertation, I pose a more general question in the final study: how scientific practices should be changed if as a field, we want to achieve more effective evidence accumulation on the key effects influencing decision-making and behavioral interventions building on them.

Factors influencing decision-making. In the last decades, hundreds of factors influencing decision processes have been explored. The paradigmatic approach of the field, the heuristics[2] and biases (HB) research program, investigated the different ways in which some of these influencing factors can cause systematical deviations from some normative standard or optimum (e.g., preference reversals, hindsight bias, framing, outcome bias, planning fallacy, sunk cost effect)[3]. Numerous competing explanations were developed on why such biases occur. Although there is no single theory which could account for the *why*s behind all of the biases, there have been some attempts in the recent years to create over-arching models providing explanations at least for some of them. One such theory is the default-interventionist theory, which assumes that many of these biases happen because individuals produce an incorrect heuristic response and fail to override this first response with further deliberation (Evans, 2007; Kahneman & Frederick, 2002; Stanovich, 2009). However, many methodological concerns have been raised (e.g., Aczel, Bago, Szollosi, Foldes, & Lukacs, 2015) and contradictory findings have been found (Thompson & Johnson, 2014) resulting in the loss of confidence in some of the building pillars of the default-interventionist theory. In Study 1 and Study 2, I examine a core pillar of the default-interventionist theory using the Cognitive

[2]It is a rarely known fact that the term heuristics was invented by a Hungarian mathematician, George Polya in his book *How to Solve it* in 1945.
[3]For a more extensive list see Baron (2008)

Reflection Task (CRT) and the denominator neglect task: how the first intuitions and the lack of intervening analytic thinking influence individuals' decisions.

Study 3 investigates a different influencing factor. It focuses on the embodied choice theory which discusses how bodily movements (action dynamics) can influence decisions. Previous experiments conducted in unconstrained spaces (Burk, Ingram, Franklin, Shadlen & Wolpert, 2014; Lepora & Pezzulo, 2015; Resulaj, Kiani, Wolpert & Shadlen, 2009) did not allow researchers to disentangle which of the two possible paths are responsible for the influence of action dynamics on decisions: do movements change the cost or the value of the response options. In Study 3, my colleagues and I aimed to design an experiment to separate the effect of these two factors.

Based on the insights of the studies examining the key effects influencing decisions, several theory-based behavioral interventions have been developed in order to support important societal goals. Such behavioral interventions aiming to influence individuals' decisions were applied by government agencies in more than 40 countries having a significant positive impact on millions of people's lives (e.g., Hansen, Skov, & Skov, 2016). However, due to methodological concerns first raised by the broader field of general behavioral interventions (e.g., Michie et al., 2016) it has been suggested that the accumulation of the scientific knowledge on when and why these interventions work is too slow. Accordingly, Study 4 had a more general scope: we reviewed the studies in the field of choice architecture interventions and proposed several ways how to improve practices leading to the improvement of the field.

Tracking decision processes. Another theme connecting the four studies of the dissertation concerns the cognitive processes of decision-making. Based on Schulte-Mecklenbeck et al. (2017), I operationally define processes "as time-dependent, pre-decisional observations which inform theories on the psychological mechanisms assumed to operate prior to decisions" (442 p.) in other words, a process is a way how a decision is made and what steps lead to it. Many theories concerning the factors influencing decision-making have process-level assumptions, including the embodied choice and default-interventionist theory. Competing theories often only differ in their process-level assumption; in what they assume happens in the "black box", between the input and the output.

However, these process level assumptions are often not tested, and researchers only examine the final decisions of the reasoner but not the processes themselves (Fiedler, 2015).

On a similar note, the exploration of processes and their integration into the decision theories are also important for practical reasons: they can increase the accuracy of the predictions about when and why different factors have an impact on decision-making and to what extent interventions build on these effects work. Accordingly, in Study 1 and Study 2, my colleagues and I applied process-tracing techniques to test previously unexamined process-level assumptions of the default-interventionist theory. In Study 3, we manipulated the processes to investigate the predictions of the embodied choice theory. In Study 4, which is a systematic scoping review, I argue why the understanding of processes and psychological mechanisms are necessary to create effective behavioral interventions.

Investigating factors influencing decision making: the default-interventionist view

The heuristics and biases research program produced a large amount of scientific knowledge on how different factors can deviate decisions from some normative standard or optimum. The default-interventionist theory is one of the most widely used theory explaining why such deviations occur. It assumes that when individuals face decisions, rapid autonomous processes generate a default response (Evans & Stanovich 2013). If this heuristic processing is followed by an intervening analytic processing, it may or may not alter the output of the heuristic process. If the first heuristic response is incorrect (biased) and the analytic reasoning fails to intervene, then the decision maker produces a biased decision.

Note that there are numerous models of decision biases with slightly different specifications and scope. However, this section only focuses on theories having two distinctive features regarding the temporal pattern behind decision biases: an initial incorrect response and the (lack of) intervening deliberative process. Following Evans (2007), I use the "default-interventionist theory" as an umbrella term to describe such models, even though it is possible that some of the authors whose theory I present didn't use this term to describe their theory.

In the following section, I focus on the description of three influential theories which I think can offer a good prototypical representation of the default-interventionist view of decision biases. This is followed by a more focused literature review leading to the motivation of Study 1 and Study 2.

Three default-interventionist theories

Theory of attribute substitution: In their inseminating work, Kahneman and Frederick (2002) aimed to provide a theory explaining many of the decision biases based on the general idea of attribute substitution. The authors suggested that when people try to solve a complicated problem, they often provide the answer to a less complicated problem without even being aware of the substitution.

A famous study from Strack, Martin and Schwarz (1988) can be used to demonstrate the theory of attribute substitution. Strack, Martin and Schwarz asked college students to report how happy they were with their life in general and also how many dates they had in the previous month. If the participants reported their general life satisfaction first and were asked about the number of dates afterward, there was a negligible correlation between the answers of the questions. However, when the questions were asked in a reversed order, the researchers found $r = 0.66$ correlation between the two variables. Kahneman and Frederick (2002) suggested that this correlation is a result of the fact that when individuals answer a complicated question, they often use peripheral, simple cues to provide the answer. In the example above, the dating question triggered an affective evaluation which was later used by the students to answer how satisfied were they with their life. That is, an automatic judgment triggered by the date question lead to a biased response to the life-satisfaction question.

The authors proposed that similar processes are responsible for many other biases. For example, in one of the classic studies, Kahneman and Tversky (1973) showed their participants the following scenario:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

After reading the scenario, participants were asked to answer what is more probable. 'Linda is a bank teller' or 'Linda is a bank teller and active in a feminist movement'. Although it is relatively easy to realize, that the probability that two events occurring in conjunction cannot be higher than the probability of either of them occurring alone[4], most of the participants

[4]If A and B are the two events, this could be formally written as $Pr(A \cap B) \leq Pr(A)$ and $Pr(A \cap B) \leq Pr(B)$.

choose the second option. Kahneman and Frederick suggested that this biased responding happens again due to attribute substitution. As the description of 'Linda' is similar to a stereotypical feminist, it automatically triggers an answer: Linda should be a feminist. Subsequently, reasoners choose the answer based on the first answer that comes to their mind without further deliberating about the logical validity of their answer (Aczel, Szollosi & Bago. 2016).

Finally, consider the bat and the ball problem, the task also described in the foreword:

A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?

Incorrectly, a high proportion of the respondents, including students from Harvard and MIT, provide '10 cents' as the answer to this task (Frederick, 2005). Kahneman and Frederick assumed that this happens because similarly to the examples above: the bat and the ball problem triggers an immediate incorrect response in most people's mind, and instead of engaging in further analytical reasoning, many of the reasoners trust the first response which comes to their mind and report this simple answer as a solution.

In sum, according to the attribute substitution model, the same pattern can be recognized across many different decision contexts such as life-satisfaction judgments, the Linda problem or the bat and the ball task: intuitive processing generates automatic responses, and deliberative processes may or may not intervene resulting in the endorsement or the correction of the first intuitive answer[5].

Finally, one important theoretical consideration has to be added: Kahneman and Frederick argued that these results and the theory of attribute substitution fit nicely the dual-process view of decision biases where cognitive processes are categorized into two main families – intuitive and analytic processes. They argued that the main feature which distinguishes these two types of processes are their speed, content, and controllability, and describe the differences in dichotomies such as automatic-controlled, effortless-effortful, associative-deductive, rapid-slow, parallel-serial, skilled action-rule application (for a critique see Keren & Schul, 2009; Gigerenzer, 2011).

[5] This model forms the basis of many of the thinking described in the influential book of Kahneman (2011), talking about the two processes as two homunculi, the lazy controller and the associative machine, sitting in the brain and being the root of many thinking biases.

Stanovich's theory of reflective mind: Stanovich (2009) proposed that monitoring should be integrated as an independent layer into the model of decision-making. Similarly to Kahneman and Frederick (2002), he also assumed that intuitive and deliberative processes, which he called Type 1 and Type 2 processes, are different. However, he suggested the model should be supplemented with a new level: the level of reflective mind.

According to Stanovich, Type 1 processes are evolutionary adaptive and closely related to ancient reproductive goals. However, he noted that fact that these processes are evolutionary adaptive doesn't mean that they are always correct on the individual level. It is reasonable to assume that the same processes which helped the hunters succeed in an ancient tribe to kill a mammoth do not necessarily help the homo sapiens make the right financial decisions, such as choosing the best investment from the hundreds of available products in a commercial bank (See also Kokis et al, 2002, Stanovich and West, 1998c, 1998d)[6]. Stanovich suggested using the acronym TASS (The Autonomous Set of System) for Type 1 processes, as Type 1 processes include a wide variety of different processes: behavioral regulations, rules, stimulus discriminations and decision-making principles which have been learned and practiced until they become automated (see also Shiffrin and Schneider, 1977). He argued that the common feature in these processes is that they are automatically triggered by the stimuli and their execution is not controlled by Type 2 processing.

In Stanovich's model, Type 2 processes are divided into the algorithmic and the reflective level of processing. The reflective level is the initiator of the algorithmic level. It is supposed to monitor the processes executed by Type 1 processes. When these automatic processes appear to be in conflict with the long-term goals, the reflective system overrides the processes of TASS and initiates the algorithmic processes. The algorithmic level becomes only active when higher-level goals states or epistemic dispositions require so and is responsible for implementing conscious calculations.

Stanovich reasoned that as most of the heuristics and biases task do not require complex computations, biased reasoning happens when the reflective mind fails to trigger the algorithmic mind. Evidence for this thinking comes from studies where intelligence or thinking dispositions were correlated with from individuals' tendency to commit several types of decision biases. According to Stanovich, intelligence measures the efficiency of the algorithmic mind while thinking dispositions provide insight on the performance of the reflective mind. It

[6] Note, that many Type 1 processes are applicable in both contexts, but this is not necessarily so.

was found that intelligence is relatively independent of decision biases but thinking dispositions significantly predict performance on a wide variety of heuristics and biases tasks (Bruine de Bruin, Parker & Fischhoff, 2007; Parker & Fischhoff, 2005; Stanovich & West, 2007; Toplak & Stanovich, 2003). That is, the performance of the reflective mind seems to predict decision biases while algorithmic mind does not.

Stanovich's theory represent a default-interventionist based tripartite model of cognitive functioning. On the most basic level, there is the TASS representing automatic processes, which produce spontaneous responses. As these Type 1 processes are autonomous and were developed to serve evolutionary purposes, they can often offer incorrect responses which can lead to biased decisions. These responses are monitored by the reflective processes which may activate the algorithmic mind leading to the override of the automatic processes.

A three-stage model of analytic thinking: Pennycook, Fugelsang and Koehler's (2015) recent model provided an explanation of why biases occur by focusing on the factors which trigger analytic thinking.[7] Their three-staged model also builds on a dual-process view of reasoning. That is, Type 1 processes are autonomous, fast and intuitive and provide default outputs which can be biased. In contrast, Type 2 processes are slow, resource demanding and reflective and may override the biased outputs of Type 1 processes. The additional question asked by Pennycook and his colleagues was the following: what are the factors which trigger analytic thinking? what makes people engage in deliberative reasoning? In line with Stanovich's reasoning (2009), they propose that Type 2 processing is not responsible for the initiation of itself. Instead, it is assumed that the conflict between the automatically triggered heuristic responses is responsible for the initiation of Type 2 analytic processes. This idea builds on several other contemporary theories, which show that decisional context may trigger several automatic responses at the same time (e.g., De Neys, 2012; Thompson, 2009; Evans, 2009). Consider, for example, the next base-rate task (De Neys & Glumicic, 2008):

In a study, 1000 people were tested. Among the participants, there were 995 nurses and 5 doctors. Paul is a randomly chosen participant of this study. Paul is 34 years old. He lives in a beautiful home in a posh suburb. He is well spoken and very interested in politics. He invests a lot of time in his career. What is most likely? (a) Paul is a nurse. (b) Paul is a doctor

[7] Using Stanovich's (2009) nomenclature, one could say that this model theorizes the factors which determine the initiation of the algorithmic mind.

As the information provided about Paul is stereotypical and strongly diagnostic of a doctor, participants are inclined to answer that Paul is a doctor (Barbey & Sloman, 2007). However, by calculating the base rate probabilities, one can conclude that the probability that Paul is a nurse is 99.5%. Recent research suggested, that people are able to process probabilities and base-rates very quickly, even by Type 1 processes, in the base-rate task (Bago & De Neys, 2017; Pennycook, Fugelsang & Koehler, 2012; Pennycook & Thompson, 2012). Pennycook and his colleagues argued that when different answer options are triggered by Type 1 processing, the conflict between these options can activate the engagement of the analytic system. According to Pennycook and his colleagues, the evidence arises from studies where people spend more time on solving problems in the conflict than in no conflict version of the same tasks (e.g., Banks & Hope, 2014; De Neys & Glumicic, 2008; Goel, Jia, & Li, 2014).

In sum, their model suggests that Type 1 processes are automatically triggered. If only one default response is triggered, people do not engage in Type 2 processing. If more than one default responses are triggered and these responses happen to be in conflict, Type 2 processes are activated and the reasoner can resolve the conflict by analytically calculating an alternative answer.

Although the three presented theories have different scopes and they use different nomenclatures, they all assume the same default-interventionist temporal structure behind biased decisions: individuals initially produce an incorrect answer and if they fail to engage in analytic reasoning[8] they provide an incorrect final answer or make a suboptimal decision. However, there are two alternative temporal-structures, mostly neglected by the default-interventions theories, which can also lead biased decisions. First, even if reasoners realize that the default response is incorrect and engage in analytic reasoning, they may fail to change the default response for not knowing the logical rules necessary to compute the correct response (see Stanovich, 2009). Second, it is also possible that the first answer, which comes to the decision makers mind, is correct, but after having a second consideration they change their mind and give an incorrect final answer or make a biased decision.

[8]Another common feature of these models is that they build on the dual-process view of human decision-making distinguishing intuitive and analytic processes. Note, that while Study 1 and Study 2 investigate the time-course assumptions of these models, they don't test whether a single or dual process view provides a more accurate description of the findings.

Study 1 and Study 2 in the present dissertation aim to investigate whether people make biased decisions because their thinking follow the temporal structure of the decision processes suggested by the default-interventionist view, or follow the alternative paths outlined above. Study 1 focuses on the Cognitive Reflection test, which is a widely used test aiming to measure individuals' capacity to suppress the first answer that comes to mind and substitute it with another correct response. In Study 2, by developing and applying a novel mouse-tracking analysis, I examine to what extent individuals' propensity not to override the first answer that comes to mind and to what extent other factors are crucial causes behind the incorrect responding on the denominator neglect task. In the following section, the literature review is motivated to present a closer look on the how Study 1 and Study 2 fit into the landscape of previous scientific findings by explaining why I chose these tasks and what previous empirical studies suggested about the processes involved. I also provide additional information about the applied process-tracing methods.

Default-interventionist theory and the Cognitive Reflection Test

The Cognitive Reflection test is the most widely used test aiming to measure people's propensity to override the first answer that comes to their mind (Frederick, 2005). The performance on the CRT correlates with a people's propensity to commit a great number of decision biases such as belief bias and syllogistic reasoning, denominator neglect, myside bias, framing (Toplak, West, & Stanovich, 2011, 2014). Low performance on the CRT is furthermore associated with higher engagement of riskier behavior (Cokely & Kelley, 2009; Frederick, 2005), more myopic intertemporal choices (Frederick, 2005) and belief in the supernatural (Gervais & Norenzayan, 2012). Demonstrating the growing popularity of the test, the paper introducing the CRT has been cited more than 2,500 times[9] (600 times only in 2017).

It is widely assumed that the temporal structure of the processes involved when people solve the CRT describe a typical default-interventionist pattern: it is thought that the CRT tasks trigger an incorrect heuristic response. If people engage in analytic reasoning then they come to the correct answer. If they fail to engage in analytic reasoning, they report the incorrect heuristic response. This view is in line with the interpretation of the CRT as "the ability or

[9] Based on Google Scholar, January 2017.

disposition to resist reporting the response that first comes to mind " (Frederick p. 35), or as of miserly processing suggested by Toplak and her colleagues (2011).

Based on this view, it is assumed that even people who provide the correct answer, they have an initial incorrect response. There are several studies providing some evidence for this view. For example, in a paper and pencil version of the task, the 10 cents response have been often found crossed out next to the correct 5 cents answer (Frederick, 2005; for further evidence see also Mata, Ferreira, & Sherman, 2013; Travers, Rolison, & Feeney, 2016). It is also assumed that those who give the incorrect heuristic answer on the CRT tasks do not engage in analytic reasoning when solving the problems. Following Frederick (2005), it is thought that even a very short amount of reflection would lead to the recognition of the thinking failure and the realization of the correct answer.

However, there are some preliminary results which raise the question whether this interpretation of the CRT is overly simplistic. There are studies which suggest that individuals with higher numerical abilities can have better number intuitions (Peters, 2012), or they may have a more sophisticated gist of numbers and operations (Reyna, Nelson, Han & Dieckmann, 2009). Consequently, it is reasonable to assume that some people may do start their thinking with a correct response or strategy in the CRT. In line with this thinking, Cokely and Kelley (2009) suggested that in contrast to the late correction mechanisms, early selection control mechanisms (Jacoby, Kelley & McElree, 1999) also contribute to the correct responding on the CRT. Furthermore, it has been found that a significant proportion of the participant failed to solve the tasks of the CRT even if they were warned to think more or their attention has been raised that those tasks are tricky (Meyer, Spunt & Frederick, 2015). This finding suggests that a moment of analytic thinking is not always enough to come to the correct answer on the tasks of the CRT.

In order to unfold the steps of the reasoning process in solving the CRT, protocol analysis was employed. This method was used in several experimental psychology and decision-making studies for studying though processes (Cokely & Kelley, 2009; Ericsson & Simon, 1980; Fox, Ericsson, & Best, 2011; Tor & Bazerman, 2003). Although it has some limitations (De Neys & Glumicic; 2008; Reisen, Hoffrage, & Mast, 2008), this method has some unique advantages: we could determine without interrupting the thought process, how the reasoners started their thinking and whether they deliberated or not after providing the first answer.

Default-interventionist theory, individual differences and the denominator neglect task

It has been shown that the way information is presented can have a biasing influence on individuals' decisions. Such biases can have important societal consequences. It has been found for example, that if risk probabilities are expressed as large ratios compared to equivalent ratios of small numbers, it increases individuals' willingness to accept health-related risk (Pinto, Martinez & Abellan, 2006). The denominator neglect task is thought to be a useful paradigm to examine such biases and became a widely used in the heuristics and biases literature. In its most basic version, participants are presented with a series of trials. In each trial, they see two ratios and are asked to choose the larger one[10]. Crucially, in half of the ratio pairs (incongruent trials), the ratio with the higher total value has the numerator with the lower value. In the other half of the ratio pairs (congruent trials), the ratio with the higher total value have the numerator with higher value. It has been repeatedly found (e.g., Stanovich & West, 2001; Bonner & Newell, 2010; Thompson & Johnson, 2014) that people have a tendency to say that the ratio with the higher numerator has higher total value, even if the total value is higher of the ratio containing the numerator with the smaller value. That is, people tend to neglect the value of the denominator.

Even though on average individuals tend to neglect the denominator, some people are more susceptible to commit this mistake and provide a biased response than others (Baron, 2008, Kahneman, 2011). Cognitive capacity, as measured by cognitive ability tests, has been shown to be an important indicator of such susceptibility. People with higher IQ scores commit less normative mistakes in the denominator neglect (Kokis et al., 2002; Stanovich & West, 2001; Thompson & Johnson, 2014) and in many other heuristics and biases tasks (e.g., Stanovich and West, 2008). According to the default-interventionist view of the denominator neglect, the reason why people are biased in this task is because they have an initial response based on the comparison of the nominators and they tend to report this initial response without engaging in further analytic reasoning (Thompson & Johnson, 2014).

Most theories explaining why individual differences arise - why susceptibility to biased responding on the denominator neglect varies between individuals - build on the default-interventionist ground. Kahneman and Frederick (2002) reasoned that the analytical processes

[10]Sometimes instead of ratios, pictures of trays containing differently colored balls are presented (e.g., Thompson & Johnson, 2014)

of individuals with higher cognitive capacity can override the incorrect heuristic response more efficiently, and this is the reason why these group of people provide more normative answers on the heuristics and biases task. The more efficient processing arises from two possible sources: more intelligent people commit fewer errors of comprehension and fewer errors of application. That is, they more often apply the appropriate logical rules and they apply these appropriate logical rules more effectively. Evans (2007) argued that the reason why higher capacity people provide more normative responses is twofold: either these individuals are more willing to engage in analytic reasoning (quantity hypothesis) or they are more efficient once they do so (quality hypothesis). Stanovich and West (2008) developed a three-layer explanation. First, some people do not possess the rules necessary to solve the problems consequently they won't be able to make a correct decision. Second, even if they possess the rule, they may fail to realize that they need to override a first heuristic response and apply the rule. Third, even if they realize the necessity to override the first heuristic response, it is possible that they do not have sufficient cognitive capacity to apply the rule correctly. Similarly, De Neys and Bonnefon (2013) suggested three different loci where individual differences in thinking and biases arise ('storage', 'monitoring' and 'inhibition'). Notably, all the presented theories assume that individual differences arise after an initial incorrect response was formulated.

The findings of several recent studies, however, raise doubt about this assumption. First, Thompson and her colleagues found that people often start their thinking with a correct response in heuristics and biases tasks (e.g., Pennycook & Thompson, 2012; Thompson & Johnson, 2014; Thompson, Prowse Turner, & Pennycook, 2011). This finding suggests that sometimes there is no incorrect heuristic response triggered. Providing further evidence, Bago and De Neys (2017) observed that people gave the correct answer on the denominator neglect task even under time-pressure or cognitive load. Thompson and Johnson (2014) also reported that the respondents' IQ correlated to a similar extent with the individuals first intuitive answer than with their final answer, suggesting that in contrast to the default-interventionist assumption, individual differences may also arise early in the reasoning process.

Based on these findings, in Study 2 we aim to investigate whether individual differences arise in line with the default-interventionist theory by testing three, not mutually-exclusive hypothesis: are higher capacity people a) more likely to start their thinking with correct initial answer b) more likely to stay with their initial answer when it is correct, and c) more likely to change their mind when their initial answer is incorrect.

We developed a novel mouse tracking analysis technique to investigate these questions. In a typical mouse tracking paradigm, participants' computer mouse movement is recorded when they choose between spatially separated response options on the screen. It is assumed that tracking computer mouse movements is valid to method explore decision-making processes (Freeman, Dale, & Farmer, 2011; Koop & Johnson, 2011; Koop, 2013; Spivey, Grosjean, & Knoblich, 2005; Travers, Rolison, & Feeney, 2016). The application of mouse tracking technique has some important advantages. First, in contrast to self-report measures, in a typical mouse tracking paradigm, participants do not know that their thought processes (mouse-movements) are tracked. In previous studies using self-reports (e.g., protocol analysis or the two-response paradigm), it is possible that individuals did not report their very first (and potentially incorrect) response in the denominator neglect because they wanted to look smarter. A second notable advantage of the new method is that it does not interrupt individuals' thought processes.

So far, I focused on what role the initial thoughts play in and how they influence decision-making and also presented an overview of the default-interventionist theory. In the next section, I describe the embodied choice theory which hypothesizes how action dynamics of thinking influence individuals' decision-making.

Investigating factors influencing decision-making: the embodied choice theory

Embodied cognition theories discuss how bodily states shape the mind, influence mental processes and decision making (Caramazza, Anzellotti, Strnad, & Lingnau; 2014). The logic of the theory follows: if a concept is activated in the mind, it also activates the constituent sensory-motor states. Conversely, if a sensory-motor state is activated (embodied) it leads to the activation of related concepts in the mind and even changes the evaluation of related concepts. The main evidence for embodied cognition comes from experiments where manipulated bodily states influenced decisions. For example, in one study, the gesture of making a fist lead to higher attributions of power to ambiguously acting targets (Schubert, 2004). Other studies showed how approach and avoidance movements (e.g., pushing versus pulling) influenced perceptions of social justice and evaluative processing (e.g., Chen & Bargh, 1999; Förster & Strack, 1998; van Prooijen, Karremans, & van Beest, 2006).

The embodied choice theory is a specific branch of the embodied cognition view. It proposes that action dynamics (movements) influence decisions and not just decisions guide

action dynamics (Lepora & Pezzulo, 2015). Consider, for example, the following situation: you are in a busy café looking for a table to sit down[11]. You spot a table on the far side and you start to approach it. However, before reaching the table you notice that there is a more pleasant table on the other side. The question arises whether your movement towards the first table influences your decision to which table to sit down. According to the embodied choice theory, the answer is yes. Decisions and actions are in a continuous loop. That is, actions influence sensory perceptions which influence the decisions, while decisions guide further actions.

There is strong evidence that people's choices are influenced by the effort needed to make a given choice and that the costs associated with the available option affect the choices[12]. For example, in the café example above, the evidence would suggest that the movement towards the first table influences the final decision by increasing cost of changing to the second table. In Resulaj, Kiani, Wolpert and Shadlen's (2009) experiment, participants were asked to make a perceptual choice in a noisy environment by moving a handle while the trajectory of the handles was tracked. The results showed a clear effect: as the cost of modifying the trajectory increased, the probability that individuals corrected their primary choice decreased. Burk, Ingram, Franklin, Shadlen and Wolpert (2014) observed qualitatively similar findings when varied the distance between the options: the more excursion of hand was needed to change one's choice the less frequently individuals changed their choice. These empirical findings provide evidence that action dynamics can influence decision-making processes by changing the cost of the options.

Lepora and Pezzulo (2015) created a formal model of how action dynamics influence decision processes. Using data from previous mouse-tracking experiments, they observed that individuals changed their mind less often as the mouse moved closer to one of the choice options. Notably, according to their model, this happened not just because the movement increased the cost of changing, but also because the movement exerted backward influence on decision processes through changing the value of the potential options. In the café example, their embodied choice model would predict that movement towards the first table decreases the chance of changing to the other table not just by increasing cost of changing but also by the increasing the value of the first table.

[11]The example is borrowed from Aczel et al. (2017).
[12]See also the power of defaults in the next section.

However, previous experiments conducted in unconstrained spaces (Resulaj et al., 2009; Lepora & Pezzulo, 2015; Burk et al., 2014) do not allow unequivocally conclude whether the increased cost of change or the increased value of one option caused by the movement itself (the continuous act of getting closer) influenced individuals' decisions. In designs applying unconstrained spaces, it is possible that the bodily movement not, but only the increased cost of the movement had an influential effect on decision-making. In Study 3, we tested whether action dynamics influence individuals' decisions by increasing the cost of change or the value of the options. In order to be able to differentiate between the effect of value and the cost, we created a novel task (Guided Movement Task, GMT). The most important feature the GMT paradigm is that it restricts the participants' movement into systematically manipulated tunnels. The fact that both the direction of the tunnels and the distance from each of the choice options are manipulated, allowed us to disentangle the effect of movement from the effect of costs and to conclude whether action dynamics influenced decision-making in our paradigm by changing the value of the options.

Up to this point, I focused on the description of two theories discussing how initial thoughts and action dynamics influence human decision-making. In the final part of the introduction, I ask a more general question: how scientific practices should be changed if as a field, we want to have better insights and make a better use of the insights produced by the studies investigating the key factors which influence decision-making.

Improving scientific practices

One of the great promises of discovering the key factors influencing decision-making is that this knowledge can be translated into useful actions: guidelines and recommendations can be created and used by individuals and policymakers to design decision environments which facilitate better decisions. This promise was popularized by Thaler and Sunstein's book, the 'Nudge' (2008). From a societal perspective, the so-called nudging (or choice architecture) approach[13] has three important defining characteristics: it aims to advance the public good, the

[13]Note, that there are several other approaches aiming to improve human decision-making. While the choice architecture movement put the emphasis on how decisions can be improved by changing the choice environment, others rather emphasized how the decision maker can be 'modified' ('taught', 'debiased'). For a more general overview see Soll, Milkman, Payne (2016).

decision makers are able to freely make any decisions and there are no costly economic incentives used.

Nudge researchers typically use important insights from the psychology literature to create behavioral interventions. These behavioral interventions are used to influence individuals' decisions. For instance, one important insight nudge interventions build on is that individuals have limited cognitive resources (Simon, 1956; Burk et al. 2014). Consider the case of organ donations. It was shown that 85% of the US population supported organ donation, but only 28% of the population made an active decision to sign a donor card. As a result, there were over 45 000 people waiting for organ donations in the United States. Similar patterns were also found in Germany, Spain or Sweden (Johnson & Goldstein, 2003). However, in several countries, including Hungary and Austria, the effective consent percentage was over 99%. Johnson and Goldstein (2003) brought evidence, that the differences between these groups of countries were not driven by cultural differences[14], but rather by the differences in the default choice offered by these countries. In Spain, USA and Germany, people need to actively offer their organs to make possible an organ transplant after their death (opt-in default). However, if Austrian or Hungarian citizens do not actively make a decision otherwise, their organs can be used to save another people's life (opt-out default). Johnson and Goldstein argued that this effect is driven by the cognitive laziness of the decision makers. People often stay with the default option, even if it goes against their preferences[15]. Whether a country applies an opt-in or opt-out default can be considered a choice architectural tool as it improves public good without using regulatory processes and economic incentives.

There are numerous nudge interventions techniques such as reframing, making information visible, providing social reference point, changing option related effort, changing range or composition of options, change option consequences, providing reminders or facilitate commitment which have been successfully applied across different contexts (Münscher, Vetter & Scheuerle, 2015). Due to their low costs and apparent success, nudge interventions seem to provide easily implementable and generally applicable tool for decision and policy makers. However, systematic reviews investigating specific nudge interventions raised several concerns regarding the amount and quality of the accumulated scientific knowledge. For instance, Skov,

[14] Consider for example Germany and Austria, which have arguably similar culture and the same language, but the organ donation rate differs by a great magnitude; 12% and 99%, respectively (Johnson & Goldstein, 2003).
[15] Note the similarity with the default-intervention theory.

Lourenco, Hansen, Mikkelsen, and Schofield (2013) reviewed the studies on choice architecture interventions aiming to change eating-related behavior. The authors observed that the general quality of the studies was low, the description of the outcome measures was inaccurate and the sample sizes were often very small. Other systematic reviews came to similar conclusions (Hollands et al., 2013; Nørnberg, Houlby, Skov, & Peréz-Cueto, 2015; Wilson, Buckley, Buckley, & Bogomolova, 2016). Altogether, the reviews suggested that at least on the investigated subdomains, decision makers had access to the tools (behavioral interventions) to influence human behavior, but their knowledge was very limited to predict when and why these tools work and to say which tool they can use in a given situation[16].

In Study 4, we intend to figure out how the field of nudging could more effectively accumulate knowledge on when and why nudges work and which nudge to use in a given situation to reach important policy goals. In order to identify the most important obstacles that can hinder the evidence accumulation of the field, we provide an overview of the characteristics of the interventions and experimental designs employed in the nudge literature. Additionally, we aim to reveal the most detrimental scientific practices and convey recommendations for future studies.

[16] Governmental behavioral science units also typically test and compare the effectiveness of several behavioral interventions in the given context. Consider for example a case from the United Kingdom, where large scale randomized control trial was conducted to promote organ donation. The researchers designed eight behaviorally informed messages (e.g., using reciprocity, social norms, loss frame, gain frame) and sent out these different versions to more than one million people in the UK. The reciprocity message ('If you needed an organ transplant would you have one? If so please help others.') was found to be the most successful increasing organ donation by more than 3%. According to the calculations this kind of messaging increased the number of organ donors around 100 000 per year (Behavioural Insights Ltd, 2015).

The Cognitive Reflection Test Revisited: Exploring the Ways Individuals Solve the Test (Study 1)

Szaszi, B.[1,2], Szollosi, A.[3,1], Palfi, B.[1], Aczel, B.[1]

[1]Institute of Psychology, Eötvös Loránd University, Budapest, Hungary
[2]Doctoral School of Psychology, Eötvös Loránd University, Budapest, Hungary
[3]School of Psychology, The University of New South Wales, Sydney, Australia

Abstract

Individuals' propensity not to override the first answer that comes to mind is thought to be a crucial cause behind many failures in reasoning. In the present study, we aimed to explore the strategies used and the abilities employed when individuals solve the Cognitive Reflection Test (CRT), the most widely used measure of this tendency. Alongside individual differences measures, protocol analysis was employed to unfold the steps of the reasoning process in solving the CRT. This exploration revealed that there are several ways people solve or fail the test. Importantly, 77% of the cases in which reasoners gave the correct final answer in our protocol analysis, they started their response with the correct answer or with a line of thought which led to the correct answer. We also found that 39% of the incorrect responders reflected on their first response. The findings indicate that the suppression of the first answer may not be the only crucial feature of reflectivity in the CRT and that the lack of relevant knowledge is a prominent cause of the reasoning errors. Additionally, we confirmed that the CRT is a multi-faceted construct: both numeracy and reflectivity account for performance. The results can help to better apprehend the "whys and whens" of the decision errors in heuristics and biases tasks and to further refine existing explanatory models.

Introduction

In the decades-long aim of psychological research to understand errors in human thinking, the Cognitive Reflection Test (CRT; Frederick, 2005) has become a pivotal tool to measure a unique dimension of individual differences. The three-item test was originally created to assess one type of cognitive ability or disposition: the capacity to suppress the 'incorrect intuitive' answer and substitute it with the correct one[17]. The *bat and the ball problem* is the most well-known example from the test: *A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?* The task can trigger a misleading answer (in this case, 10 cents), which the participants need to overcome before engaging in further reflection to arrive at the correct solution (5 cents). These supposed steps of the reasoning process make the CRT a paradigmatic demonstration of the fallibility of human thinking.

Since its publication, the original paper introducing the CRT (Frederick, 2005) has been cited over 1,900 times[18]. The cause of its popularity is multifaceted: it possesses high face validity, it is easy to administer, it predicts decision performance in many different situations, and it correlates with a great number of other measures. Just to highlight a few examples, individuals with higher CRT scores are more disposed to avoid decision biases (Toplak, West, & Stanovich, 2011, 2014) and perform better on general ability measures (Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Stupple, Ball, & Ellis, 2013). The CRT also predicts intertemporal behavior (Frederick, 2005), risky choice (Cokely & Kelley, 2009; Frederick, 2005), utilitarian moral judgment (Paxton, Ungar, & Greene, 2012), conservatism (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012), and belief in the supernatural (Gervais & Norenzayan, 2012). Extended versions of the CRT have been created (e.g., Baron, Scott, Fincher, & Metz, 2014; Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2015; Thomson & Oppenheimer, 2016; Toplak et al., 2014), as the original three items of the CRT became increasingly well known to the public.

[17] The responses in the CRT are often grouped into three categories: 'intuitive incorrect' (10 cents, 100 machines, 24 days); 'non-intuitive incorrect' (any other answer); and 'non-intuitive correct' (5 cents, 5 machines, 47 days).
[18] Based on Google Scholar, Jan 2017.

Besides its growing popularity in empirical studies, the theoretical foundations of the test have been repeatedly questioned. Two closely related sets of issues prevail in the current discussions: First, what does the CRT measure? And second, what are the steps of the reasoning process when people try to solve the test?

Regarding the first issue, most researchers argue that the CRT assesses reflectivity. Two views dominate the literature about the interpretation of reflectivity. The most popular interpretation was proposed by Frederick (2005), conceptualizing cognitive reflection as *"the ability or disposition to resist reporting the response that first comes to mind"* (p35). This approach of reflectivity has been promoted by, among others, Toplak et al. (2011) who considered the CRT as a measure of *miserly processing*, referring to people's tendency to rely on heuristic processing instead of using more cognitively expensive analytical processes. The explanation of both of these research groups builds on the assumption that the key property of the CRT is that first an 'incorrect intuitive' answer comes to the mind, and then late suppression mechanisms need to intervene and override the heuristic answer to be able to reach a normative solution by further deliberation.

Cokely and Kelley (2009) were the first to extend the dominant theoretical framework that only emphasized the role of late suppression mechanisms. They argued that early selection control mechanisms (Jacoby, Kelley, & McElree, 1999) may play an important role in the reflective behavior. They proposed that people scoring higher on the CRT process information more elaborately and tend to use more thorough search processes. Baron, Scott, Fincher and Metz (2014) provided evidence for this hypothesis. In their study, they created no-lure versions of the CRT[19] and found that these items loaded on the same factor as the standard CRT items. Additionally, both types of items (lure, no-lure) correlated to a similar extent with other measures, such as the Actively Open-Minded Thinking (AOT; Baron, 1993) or belief bias syllogisms (Evans, Barston, & Pollard, 1983). As the authors did not find evidence to support the claim that the suppression of an initial response tendency is relevant in the CRT, but observed that the test assesses the extensiveness of search, they concluded that the CRT is a measure of reflection-impulsivity (RI; Kagan, Rosman, Day, Albert, & Phillips, 1964). This, in

[19] No lure CRT tasks are CRT-like arithmetic problems that supposedly do not trigger an 'intuitive incorrect' response. For example: "If it takes 1 nurse 5 min to measure the blood pressure of 6 patients, how many minutes would it take 100 nurses to measure the blood pressure of 300 patients?" (Baron, Scott, Fincher, & Metz, 2014)

turn, is an indicator of cognitive style where there is a relative preference for impulsivity (speed) versus reflection (accuracy).

There is a parallel discussion concerning the CRT as a measurement tool. It has been argued that the CRT measures solely numeracy[20] as its items are numerical tasks. Moderate to strong correlations have been found between the CRT and other assessments of numeracy (Finucane & Gullion, 2010; Liberali et al., 2012). Welsh, Burns and Delfabbro (2013) observed that the CRT has predictive power only on those heuristics and biases tasks where numeracy plays a role in arriving at the correct solution. They concluded that the CRT assesses numerical abilities rather than the inhibition of a prepotent response. Other studies, employing factor analysis techniques, found that the CRT items loaded on the same factor as other numerical items (Baron et al., 2014; Låg, Bauger, Lindberg, & Friborg, 2014; Study 1 in Liberali et al., 2012; Weller et al., 2013). Sinayev and Peters (2015) studied whether numeric abilities or cognitive reflection are responsible for the predictive power of the CRT. Based on the observed performance on the CRT, they estimated two variables: the numerical score was calculated as the proportion of correct responses, while the cognitive reflection score was computed as the proportion of 'non-intuitive' answers. They observed that only the numerical scores in the CRT accounted for performance on other decision making and heuristics and biases tasks.

However, other results support the idea that in addition to numeracy, reflective ability is also involved in solving the CRT successfully. In contrast to Welsh et al.'s (2013) findings, Campitelli and Labolita (2010) observed that the CRT correlates with tasks without mathematical component. Pennycook and Ross (2016) reviewed evidence that the CRT was predictive of a diverse range of variables even after controlling for numeracy. Liberali et al. (2012) found that the bivariate correlations between the CRT and the numeracy scales were not high and the CRT items loaded on a numeracy-independent factor based on the results of the factor analysis. The authors concluded that the CRT is not just another test of numeracy, but also added that the CRT and objective numeracy are, in fact, related. Campitelli and Gerrans (2014) applied a mathematical modeling approach to tackle the conundrum. They estimated an inhibition parameter employing belief bias syllogisms and the AOT. They also assessed a numerical parameter using a numeracy scale. The results indicated that the models including

[20] Numeracy is one's ability to store, represent and process mathematical operations (Peters, 2012).

both an inhibition parameter and a mathematical component fitted the data better than a model including only a mathematical parameter.

Most studies using the CRT employed some explicit or tacit assumptions about the steps involved in the reasoning process of the CRT. Although a few studies tried to explore these assumptions, the analyses were based on aggregated data (e.g., Mata, Ferreira, & Sherman, 2013; Travers, Rolison, & Feeney, 2016), giving rise to methodological limitations. More specifically, data aggregation can overshadow the existence of subgroups that may follow different strategies when solving the test (Fific, 2014).

According to the most common understanding of the CRT, suppression of a first answer is a necessary step for good performance. This view about the task relies on two important assumptions. First, it assumes that even those who give the correct answer start their thinking with an 'incorrect intuitive' response, although they are able to suppress it. Frederick (2005) postulated that even the correct responders consider first the incorrect answer, based on the observation that the '10 cents' answer was often crossed out next to the '5 cents' answer in the bat and ball problem. Mata, Ferreira and Sherman (2013) found evidence that a majority of the correct responders were aware of the 'intuitive response'. Nevertheless, the authors did not control in their study for the time-course assumption of the reasoning process which is theoretically crucial, as it is possible that those who indicated awareness of the 'intuitive response' may have had a correct first response and only later, during the deliberation period, did they take into account the incorrect alternative response. Travers et al. (2016) used a computer-mouse tracking paradigm, where participants were asked to choose an answer on each CRT task by clicking on one of four response options on the screen. The authors observed that individuals who solved the tasks correctly tended to move the mouse more slowly away from the 'incorrect intuitive' response than from other 'non-intuitive incorrect' response options before clicking on the correct answer. Nevertheless, based on these findings, it is difficult to conclude whether or not there were responders whose first answer was correct. The results imply only that on average, correct responders are more likely to start their thinking with the 'intuitive incorrect' response than with other incorrect answers, and not that they never start their thinking with the correct response. Furthermore, the results of some recent studies suggest that there are individuals with correct intuitions. For example, Peters (2012) argues that people with higher numeracy "rely on their superior number intuitions" (p. 32) and based on the Fuzzy Trace theory (Reyna, Nelson, Han, & Dieckmann, 2009) she also claims that they may "derive

a richer gist from numbers" (Peters, 2012, p. 32). Supporting this idea, Thompson and Johnson (2014) reported that some individuals responded normatively on reasoning tasks when they were asked to report the initial answer that comes to mind. These tasks - similarly to the CRT – are thought to trigger an incorrect response that needs to be suppressed in order to arrive at the correct answer. The authors argued that cognitive capacity drove the production of the initial correct response. Svedholm-Hakkinen's (2015) experiments provided more evidence for the same idea: when solving belief bias syllogisms, high ability people did not show the sign of belief-inhibition; that is, they seemed to start to think using normative logic.

According to the second underlying assumption of the suppression-focused interpretation of the CRT, those who give the incorrect heuristic answer do not reflect on it. Otherwise, as Frederick (2005, p. 27) argues, "even a moment" of reflection would lead to the recognition of the failure. Previous studies have found that people spend more time (Johnson, Tubau, & De Neys, 2016) and show longer distances travelled by the mouse cursor (Travers et al., 2016) on correct responses than on the 'intuitive incorrect' answers. However, these results only support the idea that, on average, people deliberate more before producing the correct responses and one cannot conclude that the incorrect responders did not reflect. Furthermore, the fact that incorrect responders were not aware of the correct response (Mata et al., 2013; Travers et al., 2016) does not imply that these individuals did not reason analytically (Elqayam & Evans, 2011). In contrast to this assumption, Meyer, Spunt and Frederick (2015) observed that many of their participants failed to solve the bat and the ball problem despite the fact that they had been warned to think carefully about it. Moreover, previous findings have also brought evidence that deliberation does not necessarily lead to the change of the initial incorrect intuition: for instance, it has been repeatedly shown that people use reflective reasoning to rationalize or justify their first thoughts in the Wason selection task (Wason & Evans, 1975; Evans, 1996; Evans & Ball, 2010).

The current research

Our study includes both exploratory and confirmatory research. First, we aimed to explore the skills required to solve the CRT successfully. To identify the crucial individual differences behind good performance on the CRT, we used one numeracy and four reflectivity tests. The rationale for using several measures of reflectivity is that there are competing theoretical concepts of reflectivity and there is no agreement on a single and valid assessment approach. Consequently, one of our aim was to find which reflectivity measure predicts best

the performance on the CRT since this analysis can help us reveal which conceptualization of reflectivity is captured by the CRT.

Furthermore, we aimed to explore the strategies employed when individuals solve the CRT. Here, we focused on two crucial questions concerning the above-detailed assumptions of the most widely used interpretation of the CRT. First, we aimed to explore the proportion of correct responses in the CRT in which the reasoners start their response with the correct answer or with a line of thought which led to the correct answer. Second, we studied the proportion of the incorrect responses in which the reasoners reflect on the answer that first comes to their mind. Note that the first and second questions focus on the correct and incorrect cases, respectively. To investigate the strategies employed, we used protocol analysis (Ericsson & Simon, 1980), which has been found to be a valid method for studying thought processes without altering performance (Fox, Ericsson, & Best, 2011; for limitations see: De Neys & Glumicic, 2008; Reisen, Hoffrage, & Mast, 2008). Besides the fact that this method has been used in several studies in the decision-making literature to track thinking processes (e.g., Brandstätter & Gussmack, 2013; Cokely & Kelley, 2009; Tor & Bazerman, 2003), we used protocol analysis as it provided some unique advantages. For instance, with the use of this method, we could differentiate individuals who deliberated after reporting a first answer from those who did not deliberate, without interrupting the reasoning process, and while still being able to keep the CRT tasks open ended and not reducing the number of alternative answer options.

We formulated a number of additional hypotheses to test the validity of the findings of the protocol analysis. First, we hypothesized that it takes more time to solve the problems correctly in cases where the responders start their response with the incorrect answer or with a line of thought leadingto the incorrect answer (‘Incorrect-start’) than when they start their response with the correct answer or with a line of thought leading to the correct answer (‘Correct-start’). Second, we expected that there would be no significant difference in terms of reaction time and social desirability between the ‘Correct-start’ and ‘Incorrect-start’ cases. Finding that individuals in the ‘Correct-start’ cases have longer reaction times or are more socially desirable would indicate the presence of a confound in our data: that is, ‘Correct-start’ people may also suppress their first thought but do not verbalize it in our protocol analysis. Third, we expected that incorrect responders spend more time on solving the problems when

they reflect on the first answer that comes to their mind ('Reflective') compared to when they do not deliberate on it ('Non-reflective').

Finally, based on the assumption that individual differences can predict the usage of different reasoning strategies (e.g., Peters, 2012; Thompson & Johnson, 2014), we aimed to test two confirmatory hypotheses. First, we hypothesized that individuals with higher numeracy scores more often have 'Correct-start' than their less numerate counterparts. Second, we hypothesized that individuals who score higher on the reflectivity scale will more often deliberate after the first answer that comes to their mind than people who score lower on the same scale. Prior to data collection, the decision was made that for the purpose of testing the hypothesis about reflectivity and deliberation we would use the reflectivity scale that had been found to best predict the CRT performance.

Methods

Participants

219 students (75% female, M = 22.04 years, SD = 2.28) participated in our study. The participants were recruited through the university subject pool and they received course credit in exchange for their participation. All participants were native speakers of Hungarian and signed an informed ethical consent form. As nine participants indicated after the protocol analysis that they were familiar with the CRT questions, they were excluded from the online session and the analysis.

Procedure

The study consisted of an offline and an online session. For the offline session, participants were invited to the lab to participate in a personal interview. First, they were informed that the session would be recorded and later analyzed. This was followed by the detailed verbal instruction of the protocol and a warm-up session which strictly followed the procedure of Ericsson and Kirk (2011). After that, participants were asked to solve the three items of the CRT[21] in the standard order whilst thinking aloud. Not to have any undesired influence, the experimenter was seated behind the participants and provided no feedback regarding the participant's performance on the CRT. Participants were asked to read aloud the

[21] The European version of the bat and ball problem was administered where the cost of the bat and the ball is given in €.

tasks, and then to think aloud while working on the questions but not to explain their thoughts. They were also requested to indicate when they felt that they are finished with the problems. Finally, participants were asked whether they were familiar with the CRT tasks.

During the online sessions, participants completed the following questionnaires and ability measures in a fixed order using the Qualtrics survey software tool in installments: AOT (Baron, 1993), Rational-Experiential Inventory (REI; Pacini & Epstein, 1999), belief bias syllogisms (BBS; De Neys, Moyens, & Vansteenwegen, 2010), Berlin Numeracy Test (BNT; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012), Semantic illusions (SI; Mata, Schubert, & Ferreira, 2014) and finally the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1991).

Materials

Numeracy Measure

We used the computer adaptive version of the BNT (Cokely et al., 2012) to measure numeracy. The BNT predicts the comprehension of everyday risk, and the performance on the CRT and many other decision-making tests more strongly than other numerical instruments. Additionally, it is able to differentiate between highly-educated individuals. The test consists of two or three questions adaptively selected based on the former answers.

Reflectivity Measures

Participants were asked to fill out the AOT (see Appendix A.1) which measures people's tendency to consider several possible answers when facing a question, to search for evidence supporting an answer other than their previously established answer, and to seek evidence against their favored answer (Baron, 1993). We used the 8-item version of the AOT (Haran, Ritov, & Mellers, 2013) supplemented by three additional items which increase the overall reliability of the original scale (Baron, personal communication).

We also administered the 20-item Rationality scale from the Rational-Experiential Inventory (Pacini & Epstein, 1999) which measures the degree to which a person engages in and enjoys effortful cognitive activity. The inventory separates the construct of Rationality from Faith in Intuition. In this test, participants are asked to indicate on a 5-point Likert-scale how much statements such as "I enjoy intellectual challenges" are judged to be true for themselves.

Three valid and three invalid BBS were presented in a random order (see Appendix A.1). Four of our items were adopted from De Neys et al.'s (2010) study, and two additional items were developed by our research group. BBS can be used as a reflectivity measure because

the supposed underlying mechanism behind performance on BBS items is the same as behind the CRT items. People tend to decide upon the logical validity of the syllogisms based on the believability of the conclusion, which is thought to be an intuitive response. Supposedly, people have to suppress the first intuition and engage in effortful reasoning to arrive at the correct answer (Evans, 2003).

A set of SI (Mata et al., 2014) were also administered. SI tests are usually used to measure the degree to which individuals process verbal or written information carefully and accurately without containing any mathematical content (Barton & Sanford, 1993; Erickson & Mattson, 1981). Consequently, we presumed that SI could potentially assess reflective processing without measuring numeracy. The SI block consisted of 6 questions containing semantic illusions where to give the right answer participants needed to realize the semantic inconsistency embedded into the question (e.g., "How many animals of each kind did *Moses* take on the Ark?") and two simple general knowledge questions (see Appendix A.1). These latter general knowledge questions were used so participants would not become suspicious once they detected the illusions. The semantic illusions were adapted from Mata et al. (2014). Based on a similar thinking, Thomson and Oppenheimer (2016) also created an alternate form of the CRT using tasks with non-numerical content.

Social Desirability Measure

Participants were also asked to fill out the BIDR (Paulhus, 1991). BIDR measures the responder's tendency to answer in a way that makes them socially desirable in order to manage self-presentation. The BIDR consists of two subscales (Self-Deceptive Enhancement, Impression Management), from which only the second one was administered for the purpose of this study. The Subscale consists of 20 items, such as "I sometimes drive faster than the speed limit", and the responders had to report their answer on a 7-point rating scale.

Bayes Factor

As no scientific inference can be made to the hypotheses from statistically non-significant results alone (Dienes, 2014), we calculated Bayes factors (B) to supplement the frequentist analyses and used it to determine whether the null results in this study imply data-insensitivity or provide evidence for the null hypotheses. B is a statistical measure which can be used to assess the degree to which the data support one hypothesis compared to another one. To interpret the B values, we employed Jeffreys's (1961) sensitivity criterion. Accordingly, B values less than 1/3 indicate substantial evidence for the null while B values more than 3

indicate substantial evidence for the alternative hypothesis. B values between 1/3 and 3 show that the data are insensitive and should not be used as scientific evidence towards any of the hypotheses. For the B calculations, we applied the B calculator of Dienes (2008) implemented in R.[22].

Results

Descriptive results of the CRT

As the first step of our analysis, we compared the descriptive results of the protocol analysis with the most commonly reported descriptive patterns from previous studies of the CRT. The data showed acceptable reliability as measured by Cronbach-alpha (0.64), which is comparable with the results of previously reported studies (Campitelli & Gerrans, 2014; Liberali et al., 2012; Primi et al., 2015; Weller et al., 2013). While, in total, 28% of the responses were correct, the participants reported the 'intuitive incorrect' answers and other incorrect answers in 60% and in 8% of the cases respectively and gave up on solving the problems in 4% of the cases. The proportion of different types of answers showed considerable variance across the tasks of the CRT. Table 1. provides a summary of these findings. Both the solution rates and the proportion of different types of answers were in line with the previous findings in the literature (e.g., Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2015). Our results were also consistent with previous results regarding gender differences in the CRT performance (e.g., Frederick, 2005): the Mann-Whitney test indicated that males scored higher (*Mdn* = 1) on the CRT than females (*Mdn* = 0), $W = 5206$, $p = .003$.

[22] In order to compute B, one has to model the predictions of the tested hypotheses. Since all of the hypotheses in the current study had directional predictions, following Dienes's recommendations (2011; 2014), we modeled the alternative hypotheses with half-normal distributions with 0 probability for negative values. We applied two ways to determine the SD of the half-normal distributions. If we had information on the effect size of the alternative model then we used it as the SD of the half-normal distribution. Otherwise we estimated the maximum possible effect size of the alternative hypothesis and we applied the half of it as the SD of the half-normal distribution.

Table 1

The Number and the Proportion of Answers per Answer-Type

	Correct answers	Intuitive incorrect answers	Other incorrect answers	Gave up
CRT1	44 (21%)	150 (71%)	5 (2%)	11 (5%)
CRT2	46 (22%)	130 (62%)	28 (13%)	6 (3%)
CRT3	87 (41%)	98 (47%)	18 (9%)	7 (3%)
In Total	**177 (28%)**	**387 (60%)**	**51 (8%)**	**24 (4%)**

Individual differences measures and the CRT performance

The first part of the follow-up online survey containing the numeracy and reflectivity measures was returned by 206 out of the 210 participants while 195 individuals (93%) completed the second survey comprising the social desirability scale. Appendix A.2 provides on overview of the descriptive statistics of the used tests. Each analysis was run with all of the data available for that test. BNT showed significant correlation with the CRT performance, r = .49, p < .001, and all the reflectivity measures (REI, AOT, SI, and BBS) also correlated significantly with the CRT (Table 2). However, after controlling for BNT, the partial correlation analysis showed that only REI, $r(178) = .26$, $p < 0.001$, and AOT, $r(178) = .20$, $p = .007$, retained a significant relation with the CRT (SI., $r(178) = .03$, $p < 0.71$; BBS, $r(178) = .11$, $p < 0.13$).

As a next step, we aimed to investigate the individual differences behind good performance on the CRT. To do that, we built standard multiple regression models to assess the variables predictive ability on the CRT performance. Firstly, all the independent variables were entered into the model, then all the statistically non-significant predictors were removed. Our final model, comprising BNT, b = .39, 95% CI [0.29, 0.48], t = 8.22, p < .001, and REI, b = 0.02, 95% CI [0.01, 0.03], t = 4.16, p < .001, fitted the data best, $F(2,203) = 48.09$, p < .001, *adj.* R^2 = .32.[23]

[23] The assumptions of the multiple regression were not met. A bootstrapping estimation of 10000 samples confirmed the results of the regression analysis.

Table 2

Correlations of the Main Variables

	Berlin Numeracy Test	Rational-Experiential Inventory	Actively Open-Minded Thinking	Semantic Illusions	Belief Bias Syllogisms
CRT	0.494**	0.291**	0.256**	0.187**	0.292**
Berlin Numeracy Test		0.143*	0.24**	0.286**	0.384**
Rational-Experiential Inventory			0.339**	0.095	0.165*
Actively Open-Minded Thinking				0.206**	0.242**
Semantic Illusions					0.224**

* $p < .05$; ** $p < .01$

Protocol analysis: exploring the ways individuals solve the CRT

Two raters, blind to our hypotheses, categorized the verbal reports using the following coding system (Table 3). First, the answer of every individual on each CRT task has been signed as correct or incorrect. Then a different categorization procedure was applied for the correct and for the incorrect answers. The coding system is summarized in Table 3 with some prototypical examples from the bat and ball problem. The result of the categorization procedure showed high inter-rater reliability, kappa = .83.

The correct answers were classified into the 'Correct start', or the 'Incorrect start' categories. All the cases where participants started their response with a line of thought which led to the correct answer (i.e., after reading the task, expressed a coherent sequence of mental steps that led her to the correct answer), or after reading a question immediately gave the correct answer, were categorized as 'Correct start'. Otherwise, the cases where the participants started their response with an incorrect answer or with a line of thought which led to an incorrect answer, but later realized their failure, were labeled as 'Incorrect start'.

The incorrect responses were grouped as 'Reflective' or as 'Non-reflective'. Regarding the incorrect cases, the categorization procedure focused on whether the participant reflected or not after reporting a first answer. A case was classified as 'Non-reflective', if the participant accepted the first answer that came to her mind without any type of consideration, or simply echoed it. Otherwise (e.g., when the participant tried to reframe the problem, re-performed the original strategy, looked for alternative strategies or answers, expressed doubt), the protocol was categorized as 'Reflective'.

Table 3

Categorization of the Verbal Reports and the Number of Cases and Individuals in Each Category

Participants' final answer	Basis of the categorization	Categories	Definition of the categories	Example	N. of cases (N. of individuals)
Correct	What does the person start to say after reading out loud the task?	'Correct start'	Starting their response with the correct answer	It's 5 cents!	124 (86)
			Starting their response with a thinking leading to the correct answer	I see. This is an equation. Thus. if the ball equals to x. the bat equals to x plus 1 …	
		'Incorrect start'	Starting their response with an incorrect answer	I would say 10 cents…. But this cannot be true as it does not sum up to €1.10…	37 (34)
			Starting their response with a thinking leading to an incorrect answer	Let's see! €1.10 minus €1 is 10 cents… Wait. that's wrong! This should be solved as an equation…	
Incorrect	What does the person say after reporting a first answer?	'Reflective'	Expressing doubt and re-performing original strategy	… but I'm not sure… If together they cost €1.10. and the bat costs €1 more than the ball. the solution should be 10 cents. I'm done.	219 (136)
		'Non-reflective'	No reflection	Ok. I'm done.	142 (106)

The data of one participant partially and the data of two individuals completely were omitted, as the audio recordings of their trials were damaged. The exclusion criterion was set before the experiment was conducted. All the cases were excluded where the raters did not agree about the grouping of the protocol, to minimize the noise in the results of the protocol analysis. As a result, 76 additional cases (12%) were omitted from the subsequent analyses. The cases where the participants gave correct and incorrect answers were analyzed separately according to the corresponding hypotheses.

Analysis of the correct cases

The protocol analysis of the correct answers suggests that the participants performed a 'Correct-start' in 124 cases (77%) and showed an 'Incorrect start' pattern only in 37 cases (23%). The 'Correct start' pattern emerged as dominant for all of the CRT items (see Appendix B.1.1), however, it was most robustly expressed for the "lily pads" task. Note that the individual protocols formed the bases of the analysis.

To test the validity of this result, further analyses were conducted. First, we tested the hypothesis that the average final response time (FRT) in the 'Incorrect start' group is longer than in the 'Correct start' group. The rationale behind this thinking is that those in the 'Incorrect start' group need to perform extra mental operations compared to those who started their response with the correct answer or with a line of thought leading to the correct answer. In this study, FRT was operationalized as latency between the points at which the participants finished reading aloud the tasks and when they indicated that their final answer had been given. Log transformation was conducted to correct for the deviations from the normal distribution on FRT data. These log-transformed data were used in the comparison of several linear mixed random-effects models[24]. The base-model contained only the participants' ID as a random intercept regressed on FRT. In the second model, a random intercept was specified for each of the CRT items. As a result, the model fit increased significantly, χ^2 (1) = 15.41 $p < .001$. In the third model, group membership ('Correct start' vs. 'Incorrect start') was added as a fixed effect which significantly increased the model fit, χ^2 (1) = 52.37, $p < .001$. The analysis revealed that

[24] We used the glmer and lmer functions from the lme4 package in R for the mixed-effect analyses (Bates, Maechler, Bolker, & Walker, 2015). The corresponding t statistics reported are based on the result of Wald t tests.

the FRT was significantly higher in the 'Incorrect start' group than in the 'Correct start' group $b = 1.02$, 95% CI[0.77, 1.29] , $t(158.81) = 7.91$, $p < .001$.

For the purposes of the current study, we defined reaction time (RT) as the time interval that happened between the end of the task-reading and the onset of the formulation of the individual's answer. Assuming that any deliberative process is expressed in terms of thinking times, if people in the 'Correct start' group also started their reasoning process with an incorrect answer or with a line of thought which led to an incorrect answer and suppressed this first thought before starting to articulate their answer, their RT should be longer than the RT of the 'Incorrect start' group. This would indicate the presence of a confound in our data. To test this hypothesis, we built a linear mixed random-effect model and conducted model comparisons in the same way for RT as we did for FRT above. We found that neither the CRT items increased the fit of the model significantly, nor did the fixed effect of the group membership. Additionally, we calculated B to determine whether this null result implies data-insensitivity or provides evidence for the null hypothesis. The analysis yielded $B_{H(0,\ 1.63)} = 0.28$, indicating evidence for the null[25][26]. Thus, we found no difference in RT between the 'Incorrect start' and the 'Correct start' group.

People ranking higher on the social desirability scale may be less likely to verbalize the first answer that comes to mind in case it is incorrect. As this could result in a possible confound in our findings, we tested the hypothesis that individuals in the 'Correct start' group score higher on the BIDR than people in the 'Incorrect start' group. We compared mixed random-effect logistic regression models where the group membership was the outcome variable. First, we specified random intercepts for each participant and then for each CRT item. This latter effect did not significantly increase the fit of the model. In the last step, BIDR was stepped into the model, but we found no evidence that the groups differ in Social Desirability. The Bayesian

[25] The "H" indicates that we applied a half-normal distribution to model the predictions of the alternative hypothesis. The first number in the bracket displays the centre of the distribution, and the second indicates the SD of the distribution.

[26] We assumed that the effect size of the H1 cannot be bigger than the average RT of the group with longer RT. Consequently, the average RT in the 'Correct start' group was taken as an estimate of the maximum effect size of H1. The half of its value was employed as the SD of the model.

analysis further supported that BIDR does not predict the group membership of the participants, $B_{H(0,\ 0.45)} = 0.015$[27].

Analysis of the incorrect cases

The protocol analysis of the incorrect answers aimed to explore whether there are people who check the first answer that comes to their mind but still fail to solve the task. The data suggest that in 142 of the 361 cases (39%) people engaged in some kind of reflective behavior after reporting their first answer, while in 219 cases (61%) people accepted the first answer that they reported without any further deliberation. We observed a similar pattern for all the CRT items (see Appendix B.1.2).

Based on the definition of the 'Reflective' and 'Non-reflective' group, one would expect that FRT in the 'Non-reflective' group is shorter than in the 'Reflective' group. To test this assumption, we again compared linear mixed random-effect models. The model comparison method followed the procedure introduced above. The base-model contained random intercept for each participant. Then, random intercept was added for the CRT items, which significantly increased the fit of the model, χ^2 (1)= 13.31 $p < .001$. Finally, group membersip was added as a fixed effect. We found that the group membership variable significantly increased the fit of the model, χ^2 (1)= 91.63, $p < .001$. The analysis revealed that people in the 'Reflective' group spent significantly more time on solving the problems than people in the 'Non-reflective' group, $b = .73$, 95% CI[0.59, 0.87], $t(349.6) = 10.24$, $p < 0.001$.

Individual differences as predictors of task solution[28]

We hypothesized that more numerate individuals start their thinking with correct strategies or have correct intuitions on the CRT more often than their low numeracy counterparts. We compared mixed random-effect logistic regression models to test whether group membership ('Correct start' vs. 'Incorrect start') is predicted by BNT performance. In

[27] As there was no previous study examining the predictive power of BIDR on the CRT performance, we applied the predictive power of the BNT as a rough estimate for the maximum effect size of H1. Thus, the half of this value was employed as the SD of the model.

[28] Although we did not formulate specific hypotheses, Appendix B.2 depicts the means and standard deviations of all the individual differences measures (BNT, AOT, REI, BBS, SI, BIDR) across the different categories created in the protocol analyses.

the first model we specified a random intercept for each participant. The CRT item variable being stepped into the model as a random factor did not increase the model fit, nor did BNT performance yield a significant effect. We calculated B in order to test whether the data supported the null-hypothesis. The analysis resulted in $B_{H(0,\ 0.45)} = 0.62$, suggesting that the data obtained are not sensitive enough to permit a conclusion[29]. It has to be added that our data showed a ceiling effect on BNT among the correct responders which is not surprising taken that CRT tasks are highly difficult. Taken together, these findings do not allow us to draw any inference regarding our hypothesis.

Our last hypothesis predicted that people in the 'Reflective' group score higher on the REI scale than the members of the 'Non-reflective' group. To test this idea, we built a linear mixed random-effect logistic regression models. First, we added a random intercept for each participant, in a model with group membership as the criterion variable. Adding random intercepts for the individual CRT items did not increase the model fit significantly. Adding REI as a fixed effect predictor failed to increase model fit significantly. The result of the corresponding Bayes factor analysis indicated that the obtained data is not sensitive enough to permit a conclusion[30], $B_{H(0,\ 0.03)} = 0.80$.

Discussion

The findings of this study deepen our understanding about how people solve the CRT and about the abilities needed for its correct solution. The results suggest that there are individuals who start their response with the correct answer or with a line of thougth which led to the correct answer when solving the CRT tasks. Mata, Ferreira and Sherman (2013, Study 5) explicitly asked the participants after solving the modified version of the Bat and Ball problem whether the typically incorrect solution came to their mind while thinking about the task. As we did, they also found that correct responders had not thought of the 'intuitive response' in a noteworthy number of cases (28%)[31], which can be interpreted as the proportion of the 'Correct-

[29] The predictive power of the BNT for giving the right answer on the CRT was taken as the maximum of the expected effect size for H1, and so the half of this value was employed as the SD of the model.

[30] We took the maximum expected effect size from a model where REI predicted the accuracy of the answer for H1. The half of its value was employed as the SD of the model.

[31] Compared to our findings, the relatively low proportion of 'Correct-start' cases could have been caused by several differences between the two experimental designs. First, unlike us, the

start' individuals. Cokely and Kelley (2009), based on the findings of their protocol analysis also argued that the significance of early selection control mechanisms is underestimated in the decision literature. Altogether, these results provide empirical evidence that the early selection processes may play an important role in solving the CRT.

The finding that the majority of the correct responders started their response with the correct answer or with a line of thought which led to the correct answer raises questions regarding the usage of the CRT as a pure measurement of the ability to override the first 'intuitive response'. In addition, our correlational results further support that the late suppression mechanism may not be the only feature of reflectivity in the CRT. We have found that the REI and the AOT were the best predictors of the CRT performance above the BNT, and not the reflectivity measures which theoretically build upon the preconception of the suppression of a first 'intuitive answer' (BBS, SI). Cokely and Kelley (2009) found that the quantity of the verbalized reasoning in risky decision making tasks was related to CRT performance. Campitelli and Labollita (2010) have found that individuals who solved more CRT tasks possessed more general knowledge and used more detailed heuristic cues. Cokely, Parpart and Schooler (2009) demonstrated that more reflective individuals provided more normatively justifiable judgments in environments where multiple diagnostic cues were available, however they also relied more on heuristic processes when there was no diagnostic cue available. Additionally, Baron et al. (2014) observed that the predictive power of the CRT does not stem from the disposition to overcome an initial intuition in moral judgments. In line with previous results, our findings support the view that the definition of reflectivity – at least when it is operationalized by the CRT - should not be restricted to the description of the ability or disposition to override gut feelings, but instead a broader RI account of reflectivity should be used embracing the general preference for speed over accuracy.

Stanovich, Toplak and West (2008) suggested a general framework to understand rational thinking errors in heuristics and biases tasks. Their classification embraces two different kinds of causes that may be behind the thinking failures. The first cause is rooted in

---

authors used the modified Bat and Ball problem. Additionally, the authors did not control for the time-course assumption of the answers, which is crucial regarding our theoretical question, as it is possible that those who indicated awareness of the 'intuitive' response may have started to think with a correct strategy, and the incorrect solution came to their mind only later. Finally, their results are based on participants' self-reports after solving the task and not on verbal protocols.

the individuals' tendency to use heuristic processing mechanisms (Simon, 1956; Stanovich et al., 2008; Tversky & Kahneman, 1974). The heuristics and biases tasks are designed to trigger automatic but incorrect responses, which can lead individuals to report this incorrect answer as it is of low computational expense. The second cause is called *the mindware problem* (Perkins, 1995); it stems from the fact that individuals lack the declarative knowledge and strategic rules that are needed to solve some problems. Consequently, even when individuals put considerable mental effort into the problem-solving process, the lack of this necessary knowledge can lead to thinking failures (Stanovich et al., 2008).

The CRT is believed to assess *"people's tendency to answer questions with the first idea that comes to their mind without checking it"* (Kahneman, 2011, p. 65). Toplak et al. argued (2011, 2014) that incorrect responding on the CRT is not a result of a mindware problem, but rather that of miserly processing. In a recent review, Pennycook, Fugelsang and Koehler (2015) considered the role of cognitive abilities "rather rudimentary" (2015, p. 426). However, we found that many reasoners are not able to come to the right solution in the CRT even if they reflect on their first answer. Consequently, the mindware problem should be considered as one of the reasons people make errors on the CRT tasks. Meyer et al.'s (2015) work also supports our findings in this regard. The authors used four different kinds of manipulation to make people reflect on the bat and the ball problem and found that throughout all conditions a significant amount of people still reported an incorrect response. Their results also suggest that the tendency to fail the task can be caused either by "hopeless" (low ability) or by "careless" (high ability, low reflectivity) behavioral patterns. A recent study of Szollosi, Bago, Szaszi & Aczel (manuscript in preparation) brings further evidence to this hypothesis: their results showed that many of the participants who failed to solve the Bat and Ball problem reported that they had verified their answer, which can be interpreted as an indication of deliberative thinking. Additionally, our finding converges with others in the literature showing that a period of reflection does not necessarily produce beneficial results (Stanovich, 2008; Thompson, Turner, & Pennycook, 2011; Thompson et al., 2013). This result raises serious concerns about the usage of the CRT as a measure of cognitive miserliness and warns that whenever the CRT is used in correlational studies, researchers have to take into consideration whether the lack of miserliness or the mindware problem could have caused the effect as the failure on the CRT tasks can be caused by both.

The responses in the CRT are often grouped into 'intuitive incorrect', 'non-intuitive correct', and 'non-intuitive incorrect' categories (e.g., Pennycook, Cheyne, Koehler, & Fugelsang, 2015). More importantly, many studies make central conclusions from the hypotheses built on this classification (e.g., Böckenholt, 2012; Brosnan, Hollinworth, Antoniadou, & Lewton, 2014; Piazza & Sousa, 2013; Sinayev & Peters, 2015). Although our study did not focus on the question of whether a response was intuitive or deliberative (Evans, 2003, 2009), the results of the protocol analysis suggest that participants deliberated after articulating a first response in 39% of the trials where they reported an 'incorrect intuitive' final response. Note that we do not mean to speculate on whether the first response was generated by intuition or deliberation, but we argue that many of the reasoners engaged in some form of reflection despite eventually reporting the 'intuitive incorrect' answer. As a consequence, the classification based only on the final answer to indicate deliberative tendencies yields a contaminated measure that could lead to biased results. Our conclusion here is in line with previous research (e.g., Elqayam & Evans, 2011; Thompson, Turner & Pennycook, 2011; Thompson & Johnson, 2014): solely based on the normativity of the responses one cannot infer whether the answer was the output of Type 1 or Type 2 processes (Evans & Stanovich, 2013), or the decision maker engaged in deliberation or not. Our results indicate that before building on the conclusions of the studies using the original classification schema, more scientific examination would be needed to investigate the validity and the reliability of the intuitive/deliberative categories.

In accord with previous findings (e.g., Campitelli & Gerrans, 2014; Del Missier, Mäntylä, & Bruin, 2012; Pennycook & Ross, 2016), our results support the idea that both reflective ability and numeracy account for the performance in the CRT. Consequently, we suggest that whenever the CRT is used as a standalone individual differences measure, one should draw only careful conclusions about the reasons behind any correlations found (see also, Aczel, Bago, Szollosi, Foldes, & Lukacs, 2015), as there is no simple way to tell whether numerical abilities or the reflective disposition are causing the effect.

However, the methodological difficulty in the dissociation of numeracy and reflectivity is rooted deeper than the reliability of the tests. Those who have better numerical abilities might have richer and more accurate intuitions (Pachur & Spaar, 2015; Peters, 2012; Reyna et al., 2009; Thompson & Johnson, 2014), or use early controlled processes (Jacoby, Shimizu, Daniels, & Rhodes, 2005; Peters, 2012), which could lead them to more accurate responding

without being reflective in reflectivity tests that are based on numerical tasks. At the same time, low numeracy can lead to low scoring even for the highly reflective individuals. (See also the *mindware problem*). Similarly, in numeracy tests, high reflectivity can lead people to put more effort into the problem-solving procedure resulting in more correct responses (Ghazal, Cokely, & Garcia-Retamero, 2014), but low reflectivity can have a detrimental effect on performance[32]. As a consequence, whenever researchers aim to assess reflectivity with numerical test based assessment tools they have to be careful about the interpretation of the findings, as it is not possible to determine only by examining the accuracy measures whether numeracy or reflectivity lead to a correct/incorrect response. However, this conclusion is not specific to the numerical domain (Szaszi, 2016), but holds true for any domain-specific reflectivity test where additional thinking effort increases the probability of successful responding (for a similar argument see Baron, Badgio, & Gaskins, 1986).

Fox et al. (2011) outlined that verbal protocols "do not assure a complete record of the participants' thoughts" (2011, p. 338). Consequently, one limitation of our thinking aloud study is that we cannot exclude the possibility that some of those who apparently started their response with the correct answer or with a line of thought which led to the correct answer did not perceive any other response option. Although the reaction time measure supported the idea that the 'Correct start' group do not need to inhibit a first answer before starting to verbalize their response, there are alternative explanations that cannot be ruled out in our experimental design. First, reaction time is a valid measure to diagnose how much thinking is being done, but it is less reliable in determining how many mental operations are occurring. Additionally, one can assume that 'Correct-start' individuals are more cognitively able than people in the 'Incorrect-start' group. Taken as a whole, it is possible that 'Correct-start' people suppress their first answer and generate a new answer or strategy in the same time-frame as 'Incorrect-start' responders generate their first answer. Finally, it is possible that 'Correct start' reasoners considered the 'intuitive response' during the reading phase, and if so, our reaction time measure would not be a sensitive measure of it.

[32] Working memory (WM) differences can bring additional complexity in the equation: people with higher working memory spam are thought to be more numerate (Peters, Dieckmann, Dixon, Hibbard, & Mertz, 2007; Reyna, Nelson, Han, & Dieckmann, 2009), but they may find the cost of additional thinking lower than their low WM counterparts (Stupple, Gale, & Richmond, 2013).

It has been argued that reflectivity is a key individual differences dimension predicting rational errors in heuristics and biases tasks (e.g., Toplak et al., 2011; Stanovich et al., 2008) and in diverse everyday situations (Pennycook, Fugelsang, et al., 2015). Our study aimed to enhance our knowledge of the CRT, as it is the most widely used behavioral measure of reflectivity. In sum, we observed that there are several ways people can solve or fail the test. Importantly, some individuals started their response with the correct answer or with a line of thought which led to the correct answer, while others fail to solve the CRT tasks even when they reflect on it. Additionally, the current results suggest that the CRT test rather measures a general preference for speed over accuracy and not just individuals' ability to suppress a first 'intuitive answer'. In our view, the CRT is a useful and important measurement tool of reflectivity. However, this study raises doubts about the validity of the studies that build on the CRT as a simple measure of analytical thinking, since the use of the CRT as a standalone predictor can easily lead to the overestimation of the role of reflectivity and the underestimation of the role of numerical ability in decision performance. As the CRT tasks are pivotal examples in several dual-process models of reasoning and decision making, the implications of our findings go beyond the CRT as a measurement tool. Our implications about the processes and abilities involved in the CRT can be used to better apprehend the "whys" and "whens" (De Neys & Bonnefon, 2013) of the decision errors in heuristics and biases tasks and to further refine existing explanatory models.

Data and analysis code

The analysis code and the underlying databases are openly available at https://osf.io/vduyt/.

Appendix

Appendix A

A.1.Materials Used

A.1.1 Actively Open Minded Thinking scale

1. Allowing oneself to be convinced by an opposing argument is a sign of good character.
2. People should take into consideration evidence that goes against their beliefs.
3. People should revise their beliefs in response to new information or evidence.
4. Changing your mind is a sign of weakness.
5. Intuition is the best guide in making decisions.
6. It is important to persevere in your beliefs even when evidence is brought to bear against them.
7. One should disregard evidence that conflicts with one's established beliefs.
8. People should search actively for reasons why their beliefs might be wrong.
9. When we are faced with a new question, the first answer that occurs to us is usually best.
10. When faced with a new question, we should consider more than one possible answer before reaching a conclusion.
11. When faced with a new question, we should look for reasons why our first answer might be wrong, before deciding on an answer.

Note. The items 1-8 were published by Haran, Ritov, and Mellers (2013). The items 9-11 were provided through personal communication by Jonathan Baron. Reverse scored items: 4, 5, 6, 7, 9.

A.1.2 Semantic Illusions

1. There is a running race among A,B,C,D,E,F. If B pass the person in second place, what place is now B in.
2. Larry's father has five sons, viz. Ten, Twenty, Thirty, Forty... Guess what would be the name of the fifth?
3. How many animals of each kind did Moses take on the ark?
4. In which decade did the Beatles become the most popular American band ever?
5. In which day of September did the Twin Towers in Washington DC get attacked by Islamist terrorists?
6. A plane was flying from Germany to Barcelona. On the last leg of the journey, it developed engine trouble. Over the Pyrenees, the pilot started to lose control. The plane eventually crashed right on the border. Wreckage was equally strewn in France and Spain. Where should the survivors be buried?

Note. The items 1-2 were collected from the internet while the items 3-6 were adopted from Mata, Schubert and Ferreira's (2014) study.

A.1.3 Belief Bias Syllogisms

Invalid/believable	Valid/Unbelievable
1. All flowers need light. Roses need light. Roses are flowers.	2. All mammals can walk. Whales are mammals. Whales can walk.
3. All dogs have snouts. Labradors have snouts. Labradors are dogs.	4. All vehicles have wheels. Boats are vehicles. Boats have wheels.
5. All fruits have corns. Apples have corns. Apples are fruits.	6. All birds have wings. Cats are birds. Cats have wings.

Note. The items 1-4 were adopted from De Neys, Moyens, and Vansteenwegen (2010). The items 5-6 were developed by our research group.

A.2 Descriptive Statistics of the Tests Used in the Study

	CRT	AOT	REI	BBS	SI	BNT	BIDR
Number of people	210	206	206	206	206	206	195
Theoretical range	0-3	11-77	20-100	0-6	0-6	1-4	20-140
Range of data	0-3	39-71	27-98	0-6	0-6	1-4	47-118
Median	1	57	75	5	2	2	86
Mean	0.8	56.7	72.1	4.5	2.6	2.4	84.9
SD	1.0	6.5	13.4	1.8	1.4	1.3	15.0

Note. Actively Open-Minded Thinking (AOT) Rational-Experiential Inventory (REI), belief bias syllogisms (BBS), Semantic illusions (SI), Berlin Numeracy Test (BNT), Balanced Inventory of Desirable Responding (BIDR),

Appendix B

B.1. Protocol Analysis Results per CRT Item

B.1.1 Distribution of Final Correct Responses per CRT Item: Number of Trials in The 'Correct start' and the 'Incorrect start' Group.

Item	'Correct start' (n)	'Incorrect start' (n)	Total (n)
CRT1	24	14	38
CRT2	28	11	39
CRT3	72	12	84
CRT	**124**	**37**	**161**

B.1.2 Distribution of Final Incorrect Responses per CRT Item: Number of Trials in the Reflective and the 'Non-reflective' Group.

Item	'Non-reflective' (n)	'Reflective' (n)	Total (n)
CRT1	78	56	134
CRT2	83	43	126
CRT3	58	43	101
CRT	**219**	**142**	**361**

B.2 Means and Standard Deviations of the Individual Differences Measures Used for Each Protocol Category (Mean (SD))

	Correct start	Incorrect start	Non-reflective	Reflective	Gave up
BNT	3.14 (1.10)	3 (1.20)	2.11 (1.18)	2.06 (1.17)	1.95 (1.21)
AOT	58.45 (6.03)	58.22 (5.55)	55.76 (6.90)	56.45 (6.24)	54.36(6.75)
REI	76.60 (11.42)	77.32 (8.57)	69.30 (13.90)	71.63 (12.70)	65.45 (19.31)
BBS	5.05 (1.59)	5.03 (1.61)	4.31 (1.89)	4.27 (1.82)	4.68 (2.06)
SI	2.86 (1.23)	2.65 (1.27)	2.53 (1.46)	2.42 (1.59)	2.64 (1.33)
BIDR	83.72 (14.97)	87.29 (14.48)	85.38 (14.54)	83.6 (14.78)	89.48 (15.72)

Note. Berlin Numeracy Test (BNT), Actively Open-Minded Thinking (AOT) Rational-Experiential Inventory (REI), belief bias syllogisms (BBS), Semantic illusions (SI), Balanced Inventory of Desirable Responding (BIDR)

B.4 The Number of 'Correct start' and 'Incorrect start' cases within the Correct and Incorrect Final Responses

	'Correct-start'	'Incorrect-start'
Correct final response	124	37
Incorrect final response	1	349

Note. We run an additional protocol analysis to separate the 'Correct-start' and 'Incorrect-start' cases within the incorrect responses. Similar to Appendix B.1.1 and B.1.2., this table only shows those cases where the raters were on agreement upon the categorization of the cases. 12 cases were excluded from the 362 incorrect responses due to disagreement among the raters.

Thinking dynamics and individual differences: Mouse-tracking analysis of the denominator neglect task (Study 2)

Szaszi, B.[1,2], Palfi, B.[3], Szollosi, A.[4], P. J. Kieslich[5], Aczel, B.[2]

[1]Doctoral School of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[2]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[3]School of Psychology, Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK
[4]School of Psychology, University of New South Wales, Sydney, Australia
[5]Experimental Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany

Abstract

Most decision-making models describing individual differences in heuristics and biases tasks build on the assumption that reasoners produce a first incorrect answer in a quick, automatic way which they may or may not override later and that the advantage of high capacity reasoners arises from this late correction mechanism. To investigate this assumption, we developed a novel mouse-tracking analysis technique to capture individuals' first answers and subsequent thinking dynamics. Across two denominator neglect task experiments, we observed that individuals initially move the mouse cursor towards the correct answer option in a substantial number of cases suggesting that reasoners may not always produce an incorrect answer first. Furthermore, we observed that compared to low capacity reasoners, high capacity individuals revise their first answer more frequently if it is incorrect and make fewer changes if it is correct. However, we did not find evidence that high capacity individuals produce correct initial answers more frequently. Consistent with the predictions of previous decision-making models, these results suggest that in the denominator neglect task the capacity-normativity relationship arises after the initial response is formulated. The present work demonstrates how the analysis of mouse trajectories can be utilized to investigate individual differences in decision-making and help us better apprehend the dynamics of thinking behind decision biases.

Introduction

In the simplest form of the denominator neglect task, participants are asked to choose the larger of two ratios. The fact that individuals often base their answer on the comparison of the numerators instead of comparing the value of the ratios suggests that simple changes in the way this information is presented may influence decisions (Bonner & Newell, 2010). For example, the perceived risk of developing cancer (Yamagishi, 1997) or the willingness to accept health-related risks (Pinto, Martinez & Abellan, 2006) can increase if the risk probabilities are expressed as a ratio of large numbers compared to an equivalent ratio of small numbers. One central goal of reasoning and decision-making research is to understand why such biases occur and why some individuals are more susceptible to these biases than others (see Baron, 2008; Kahneman, 2011). Joining this endeavor, the current research aims to investigate individual differences in susceptibility to biases and the dynamics of cognitive processes underlying those individual differences.

Individual differences in cognitive capacity[33] have been shown to be a powerful predictor of normatively correct responding in a variety of heuristics and biases (HB) tasks. This capacity-normativity relationship has been found in syllogistic reasoning problems (Stanovich & West, 1998a; Stanovich & West, 2008; Svedholm-Häkkinen, 2015), framing tasks (Bruine de Bruin, Parker, & Fischhoff, 2007; Frederick, 2005; Parker & Fischhoff, 2005; Stanovich & West, 2008), base rate tasks (Stanovich & West, 1998b), belief bias (Stanovich & West, 2008) and probability matching tasks (West & Stanovich, 2003). Similarly, people with higher IQ and SAT scores give more correct answers in the denominator neglect task. (Kokis, Macpherson, Toplak, West, & Stanovich, 2002; Stanovich & West, 2001; Thompson & Johnson, 2014).

Kahneman and Frederick (2002) argued that more intelligent individuals show better performance on HB tasks because they use their deliberative processes more efficiently to override the output of the first, incorrect heuristic response. They add that this can occur via two possible paths. High ability reasoners may be more likely to learn the necessary logical

[33]Building on the literature investigating the capacity-normativity relationship in reasoning (e.g., Stanovich & West, 2001; 2008; Thompson & Johnson 2014; Thompson Pennycook, Trippas & Evans, 2017), we define cognitive capacity as a capacity measured by cognitive ability tests. Previous studies used predominantly IQ tests or Scholastic Aptitude Test (SAT) scores.

rules (commit fewer errors of comprehension) and/or they may be better able to apply the learned rules in a more effective way (commit fewer errors of application).

Stanovich and West (2008) further developed these ideas in their framework and determined three loci from where individual differences can arise. First, if the reasoner does not have the relevant declarative knowledge and procedures (mindware) available to solve an HB task, she will end up with the incorrect answer. Second, even if the reasoner has the necessary mindware available, she has to recognize the need of applying the appropriate strategy; otherwise, she will not override the heuristic response. Finally, even if the reasoner has the sufficient knowledge to solve the problem and detects the need to override the first incorrect response, she will not come to the correct answer if she does not have the sufficient cognitive capacity for the required sustained inhibition and cognitive decoupling.

De Neys and Bonnefon (2013) applied a similar partitioning of the possible causes of the individual differences in thinking biases, using the 'storage', 'monitoring' and 'inhibition' labels to refer to the different loci of individual differences (the 'whys'). The authors suggested an additional approach to organize the literature and differentiated between early and late divergence between biased and non-biased individuals (the 'whens'). Interestingly, in their framework, even in the case of early divergence, biased and non-biased reasoners start to go on a different path only after the first intuitive response has been formulated.

Evans (2007) developed two hypotheses aiming to provide explanations for the capacity-normativity relationship. According to the *quantity hypothesis*, individuals with higher cognitive ability have a higher propensity to engage in analytic reasoning which makes them more prone to override the first heuristic answer. In contrast, the *quality hypothesis* states that better performance of higher ability individuals arises because they are more likely to come to the normative solution once they are engaged in analytic reasoning.

The common aspect of these approaches is that they all assume a specific pattern that people's decision-making process follows when solving HB tasks: initially they will produce an incorrect answer, which they may or may not override at a later point[34]. Consequently, all of these models assume that the capacity-normativity relationship arises late[35] in the decision-

[34] Note that due to the focus of the present article, the literature review only summarizes models which make predictions on the temporal dynamics of individual differences in HB tasks.
[35] In the current paper, we consider any process as late that occurs after a first response was formulated.

making process. This late correction mechanism determines whether one changes her mind from the initially produced incorrect response.

However, recent studies using the two-response paradigm (Thompson, Prowse Turner, & Pennycook, 2011) challenged the assumption that people always start their thinking with an incorrect response in the HB tasks. In the two-response paradigm, people are asked to provide an initial intuitive answer (along with other measures), after which they are encouraged to take as much time as they need to rethink their response to give the correct answer (for a detailed description, see Thompson et al., 2011). Applying this paradigm to several HB tasks (such as the denominator neglect task, the base rate task, a causal reasoning task, and a categorical syllogism task), Thompson and Johnson (2014) provided evidence that people start their thinking with a correct initial response in a considerable number of cases (see for additional supporting results Pennycook & Thompson, 2012; Thompson et al., 2011). Bago and De Neys (2017) found similar results applying time-pressure and cognitive load in the two-response paradigm which further supports the idea that people often have a correct initial response. Szaszi, Szollosi, Palfi, and Aczel (2017) employed a thinking aloud procedure to investigate the thinking processes in the Cognitive Reflection Test (CRT), a popular measure showing how HB tasks can trigger an incorrect initial answer. The authors found that in 77% of the trials with correct responses, the respondents did not begin by verbalizing any consideration of the intuitive response, suggesting that they may have started their thinking already with a correct response or with a line of thought leading to the correct response when solving the tasks of the CRT.

Some recent data also question that the capacity-normativity relationship in HB tasks arises from thought processes that occur after the first response is formulated. Thompson and Johnson (2014) found that in three of the four HB tasks investigated in their study, IQ significantly correlated with the normativity of the first answer to a similar extent as with the normativity of the final response, suggesting that IQ is associated with correct first responses. Svedholm-Hakkinen (2015) found that in contrast to the less cognitively abled, the highly skilled reasoners did not show a sign of belief-inhibition (longer reaction times) in the conflict version of the belief bias syllogisms task compared to the non-conflict version of the same task. Szaszi et al. (2017) investigated whether individuals with higher cognitive capacity, as measured by the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal & Garcia-Retamero, 2012), more often start their thinking with a correct intuition or strategy in the CRT tasks. Bayes

factor analysis revealed that their data were too insensitive to draw conclusions regarding this question.

There are two possible limitations of the previously described studies investigating individuals' first answers in HB tasks. First, as the results are based on self-report measures, it cannot be ruled out that in order to look more socially desirable, individuals do not report their very first (and potentially incorrect) response, but only an answer on which they elaborated more already. Second, a critic can argue that the fact that participants have to provide the initial answer in the two-response paradigm can have an influence on the subsequent reasoning process – even if it does not alter the final answer compared to conditions without interruption (Thompson et al., 2011).

In the current research, we developed a mouse-tracking analysis technique to assess individual differences in people's decision dynamics in HB tasks. An important advantage of this method is that it does not rely on self-reports and that it does not interrupt people's decision process. In our experiments, we recorded participants' mouse movements in a computerized version of the denominator neglect task.

Our goal was twofold. First, we aimed to investigate the assumption that reasoners first produce incorrect answers in HB tasks. Accordingly, we explored the proportion of trials in which individuals moved their mouse initially towards the correct response. Second, we investigated why higher capacity people give more normative answers. Specifically, we tested three explanations: Higher capacity people (1) have a higher likelihood for initially correct answers, (2) are more likely to stay with their initial answer when it is correct, and (3) are more likely to change their mind when their initial answer is incorrect.

We chose the denominator neglect task to test these hypotheses, firstly, because as a simple two choice reasoning problem, it is an ideal candidate for mouse-tracking analysis; secondly, because previous studies demonstrated that individual differences robustly arise in this task (e.g., Kokis et al., 2002; Stanovich & West, 2001; Thompson & Johnson, 2014). Our third reason was that the denominator neglect task contains both incongruent and congruent trials which can be used to measure the sensitivity of the mouse-cursor analysis: we expected to find fewer correct initial response and more changes of minds in the incongruent trials (see Bonner & Newell, 2010; Thompson & Johnson, 2014).

The current paper contains two experiments: a mouse-tracking experiment and its replication. Since we had to make several post-hoc changes in the analysis of Experiment 1, we

replicated the study to ensure that our findings are robust. As the methods and the analyses were identical for both experiments, we report them conjointly.

Methods

Denominator neglect task

In the denominator neglect task, participants were presented with two ratios and were asked to choose the larger one. Note that in this simple version of the denominator neglect task, we displayed only the ratio pairs but not pictures of trays and the description of the task was also simplified accordingly[36]. The ratio pairs used in the current study were taken from Experiment 2 of Bonner and Newell (2010). Every ratio pair consisted of a 'small-ratio' and a 'large-ratio'. The denominator of the 'small ratio' was always 10 while the numerator was either 1, 2, or 3. For the 'large ratio', the denominator was always 100 while the numerator changed in a way that the value of the large ratio could differ from the smaller ratio within the range of -9/100 and 9/100. In 27 trials, the large ratio had higher value (congruent trials), while in the other 27 trials, the value of the small ratio was higher (incongruent trials). It has been argued that in the incongruent trials, an incorrect heuristic response is triggered based on the comparison of the numerators (e.g., Stanovich & West, 2001; Bonner & Newell, 2010; Thompson & Johnson, 2014). As a result, people (incorrectly) tend to choose the ratio with the higher numerator while neglecting the denominator. It is to be noted that we used both the incongruent and congruent trials to test the sensitivity of our analysis, but only used the incongruent trials to test our main hypotheses, since our interest in the current study was the investigation of thinking dynamics in a task where supposedly the first heuristic answer is incorrect.

Cognitive capacity measures

We administered an adaptive IQ test (Kovacs & Temesváry, 2016) which applies Raven-like matrices and was adapted to the Hungarian population[37]. In contrast to classical

[36] There are several, more complex versions of the denominator neglect task previously used in the literature. Instead of or along with the ratios, some authors present pictures of trays containing differently colored balls (e.g., Thompson & Johnson, 2014, Bonner & Newell, 2010). In some other studies, instead of using trials of ratios of different values, researchers present pictures of trays representing identical odds of winning (e.g., Epstein & Pacini, 2000).
[37] The test is available at https://mensa.hu/tesztiras/online-iq-probateszt.

paper-pencil tests, the adaptive IQ test uses a response item-bank and the items shown to each participant is determined by the individual past performance. The procedure ends when the error range of estimation is smaller than a pre-defined threshold. Participants were also asked to fill out the Berlin Numeracy Test (BNT) which measures numeracy (Cokely et al., 2012)[38].

Procedure

The experiments consisted of two sessions, an offline and an online session. For the offline session, participants were invited in groups of 15-20 to a computer test room. The computer screens had a 1440×900 px resolution and the standard mouse-sensitivity settings for Microsoft Windows 8 Enterprise were used (medium mouse speed, acceleration turned on). The experiments were built and run in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012). The mousetrap plugin for OpenSesame (Kieslich & Henninger, 2017) was used to record the *x*- and *y*-coordinates of the computer mouse-cursors every 10 ms during the trials. At the beginning of each experiment, participants provided informed consent and read the following instructions: "In the experiment, you will see ratio pairs and your task is to choose the larger value. Use the mouse cursor to indicate your decision". No information was provided about the mouse-tracking aspect of the experiment. Afterwards, participants completed four practice trials to familiarize themselves with the task. This was followed by the presentation of 54 ratio pairs in a randomized order for each participant. Participants had to click a start button in the bottom center of the screen to start a new trial (after which the mouse-cursor was automatically relocated to a predefined start position in the bottom center of the screen). In each trial, participants were presented with two ratios it the top right and left corner of the computer screen (Figure 1). The left/right position of the ratios was randomized on the trial level. Participants had 3 seconds to make their decision in each trial[39]. Aside from the mouse movements, the accuracy and response time were recorded in each trial. After the offline session, participants

[38] The BNT has two different formats. The standard format contains four questions while the computer adaptive version of the test applies two or three questions selected based on the individual performance of the participants. In Experiment 1, some participants answered the standard format while other the adaptive version. In Experiment 2, solely the adaptive version of the BNT was administered. The adaptive BNT is available at http://www.riskliteracy.org/.

[39]We applied a three-second time-pressure to motivate people to start moving the mouse cursor as early as possible. This way, we aimed to make our tool more sensitive to track the first commitments. As we did not want to draw participants' attention to their mouse movements, we did not instruct them to initiate movement as early as possible (see Scherbaum & Kieslich, in press, for a discussion of different starting procedures in mouse-tracking experiments).

were sent an email containing the information about the online session. Here, they were asked to fill out an online test package containing the cognitive capacity measures. Only participants who completed both the cognitive capacity tests and the denominator neglect task were included in the analysis.

Measuring dynamics of thinking using mouse-tracking

In a typical mouse-tracking paradigm, participants are asked to choose between two spatially separated options on the screen while the movement of their computer mouse is recorded. It is assumed that if the decision maker considers choosing one of the choice options, she moves the mouse cursor towards that option (Freeman, Dale, & Farmer, 2011; Koop & Johnson, 2011; Koop, 2013; Spivey, Grosjean, & Knoblich, 2005; Travers, Rolison, & Feeney, 2016). To assess the temporal development of participants' choice commitments, we developed a mouse-tracking analysis using the areas of interest (AOI) technique (see Palfi, Kieslich, Szaszi, Wulff, & Aczel, 2017, for a detailed description of the method and a comparison with other methods)[40]. The main idea behind this technique is that one can explore a reasoner's first and subsequent choice commitments by creating AOIs surrounding the choice options (see Figure 1A) and analyzing the order in which the AOIs were visited by the mouse cursor in each trial (for similar approaches, see Travers et al., 2016, and Gurcay & Baron, 2017).[41] In the current study, we used the reasoner's initial commitment (i.e., first AOI around one of the choice options visited by the participant's mouse cursor) as a proxy for the participant's first answer. If this first commitment differed from the individual's final answer, we classified the trial as a Change of Mind (CoM) trial (e.g., in Figure 1C). Note, that to categorize a trial as a CoM trial, it was necessary that the first and final answer differed (as is the case in Figure 1C but not in Figure 1D). This was done as we were specifically interested in changes between the

[40]One advantage of the AOI technique is that it can identify the initial answers even if reasoners make more than one choice commitment in a trial. For a discussion on why this is an important characteristic of the analysis see the Supplementary materials.

[41]We determined the size of the AOIs by applying and adjusting Freeman's (2014) maximum deviation based technique to the task setup (especially the button positions) used in the current study. The detailed calculation can be found in Supplementary Materials - Analysis code (see also Palfi et al., in preparation). Employing this method, the two AOIs overlapped slightly at the top of the screen. The common area was split symmetrically; the left part of it was allocated to the area of the left button and the right part of it to the area of the right button (see Figure 1A).

first commitment and the final response – and not in potential additional changes happening in between.

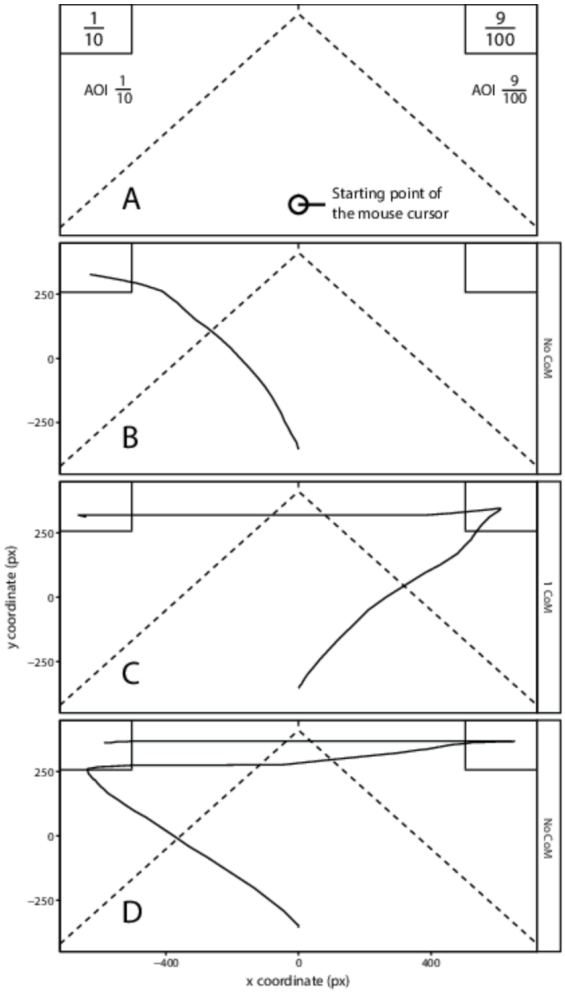

Figure 1. **Figure 1A** shows an exemplary ratio pair along with the borders of the corresponding areas of interests (AOIs) and the starting position of the mouse cursor. Figure 1B, 1C and 1D show three exemplary mouse trajectories. **Figure 1B** depicts a case where the participant moves the mouse-cursor directly to the left option. As the mouse trajectory enters only into the left AOI, here we conclude that there was only one choice commitment towards the left option and there was no CoM. **Figure 1C** shows a case where ultimately the left option is chosen but the cursor was first moved into the right AOI. Here, we conclude that there were two choice commitments, the first commitment was to the right option which differed from the final answer, consequently we classify this as a CoM trial. **Figure 1D** illustrates a case when the individual moves the mouse cursor first into the left AOI, then to the right AOI and finally again to the left AOI. Here, we conclude that the individual was first committed towards the left option, then towards the right option before finally choosing the left option. We categorize such trials as no CoM, since the first commitment and the final answer were the same.

Analysis

Analyses were performed using the statistical programming language R (R Core Team, 2013). Mouse movements were processed and analyzed using the mousetrap R package (Kieslich, Wulff, Henninger, Haslbeck, & Schulte-Mecklenbeck, 2016). In the analyses, choices and responses were predicted in linear mixed models using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015), specifying a random intercept for each participant (the tested models are available in the Supplementary Analysis Code). For response times, we used a linear mixed model and *p*-values were obtained with the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017); for dichotomous outcomes (choices, correct first answers, changes of mind), we used a generalized linear mixed model with a binomial link function.

Participants

Experiment 1: Participants were recruited from a local university subject pool in Hungary and received course credit in exchange for participation. 109 undergraduate students provided answers on the cognitive capacity tests and the denominator neglect task. The participants who provided only 0 or 1 (out of 27) correct answers in one of the conditions of the denominator neglect task (8 participants) and those participants who obtained (for university students) unrealistically low scores on the IQ test (<85, 4 participants) were excluded as these patterns indicated that the participants were unmotivated to follow the instructions. Furthermore, the trials on which participants failed to respond within the 3 second time limit were omitted from the analyses (4% of trials). 97 participants were included in the final analyses (88% female, mean age = 20.95 years, *SD* = 1.75; all Hungarian native speakers).

Experiment 2: Participants were recruited from the same local subject pool excluding those who participated in Experiment 1. From the 143 participants that completed the cognitive capacity measures and the denominator neglect task, 10 and 4 participants were excluded respectively employing the same exclusion criteria (and 4% of trials were excluded as they exceeded the 3 seconds time limit). The remaining 129 participants were included in our final analyses (53% female, mean age = 20.27 years, *SD* = 1.55; all Hungarian native speakers).

Results

Table 1 summarizes several aspects of participants' responses: accuracy of the first answer, accuracy of the final answer, CoMs, and response times in both the congruent and in

the incongruent conditions. While the accuracy of the final answer was significantly higher in congruent compared to incongruent trials, the difference was only small on a descriptive level. However, participants took significantly longer to respond in incongruent trials. The mouse-tracking method showed the expected pattern regarding the effect of congruency: there was a substantially lower percentage of correct first answers in incongruent compared to congruent trials. Likewise, there were more CoMs in incongruent compared to congruent trials. As can be seen in Table 2, all differences were statistically significant.

Table 1

Descriptive Statistics of Accuracy of the Final and the First Answer, Changes of Minds, and Response Times in Experiment 1 and 2

		Accuracy of first response (%)		Accuracy of final response (%)		Change of Mind (%)		Response time (ms)	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Experiment 1	Congruent	73	44	85	36	20	40	1479	495
	Incongruent	45	50	84	37	45	50	1625	498
Experiment 2	Congruent	69	46	85	36	26	44	1567	477
	Incongruent	43	50	85	35	49	50	1695	469

Note. Means and standard deviations are calculated based on the trial level values (ignoring participants).

Table 2

Results of (Generalized) Linear Mixed Models for Different Predicted Variables in Experiment 1 and 2 with Congruency as a Predictor

Predicted variable	Experiment	Estimate (*OR*/*b*)	95% CIs	Test statistic (*z*/*t*)	*p*
Accuracy of first response	1	3.80	[2.81, 5.14]	8.67	< .001
	2	3.30	[2.62, 4.16]	10.07	< .001
Accuracy of final response	1	2.16	[1.33, 3.54]	3.08	.002
	2	2.23	[1.28, 3.86]	2.84	.004
Changes of mind (CoM)	1	0.29	[0.23, 0.36]	-10.66	< .001
	2	0.34	[0.28, 0.41]	-11.26	< .001
Response time (RT)	1	-152.99	[-181.78, -124.20]	-10.42	< .001
	2	-129.78	[-155.67, -103.89]	-9.83	< .001

Note. The estimates represent the change in the estimate in the incongruent condition compared to the congruent condition in a (generalized) linear mixed model on the trial level specifying a random intercept and random slope per participant. For RT, unstandardized coefficient estimates (*b*) and *t*-values are reported; for all other variables, odds Ratios (*OR*) and *z*-values are reported.

Exploring the dynamics of thinking: accuracy of first and final answers

Since our interest in the current study was the investigation of thinking dynamics in a task where supposedly the first heuristic answer is incorrect, in the subsequent analyses we analyzed only the incongruent trials. Table 3 shows the number of correct and incorrect final responses in the incongruent trials based on what the participant's first commitment was. As expected, in many trials, participants started to move the mouse towards the incorrect answer first; they mostly ended up changing their mind and choosing the correct answer in the end. Still, several trials with an initially incorrect response resulted in incorrect final responses. However, there were also several trials in which individuals' first answer was the correct response and where this also corresponded to the final answer. If the initial response was correct, participants changed their mind only in a small number of cases and provided an incorrect final response.

Table 3

Percent of Trials (in the incongruent condition) per Experiment Classified Based on the Correctness of the Initial and Final Response

		Correct initial response	Incorrect initial response
Experiment 1	Correct final response	42%	42%
	Incorrect final response	3%	13%
Experiment 2	Correct final response	40%	45%
	Incorrect final response	3%	11%

Dynamics of the capacity-normativity relationship

To explore whether cognitive capacity predicts individual differences in this task, we first investigated whether higher capacity people gave more normative responses in the incongruent trials of the denominator neglect task. We calculated composite scores from the IQ and the BNT[42] scores and used these as an indicator for individuals' general cognitive capacity. These composite scores were calculated as the sum of the *z*-transformed IQ and the BNT scores. Then we regressed the accuracy of the final answer on the composite score in a generalized linear-mixed model. As expected, the composite score predicted overall normative accuracy in both experiments, as the odds of accurate answers increased with increasing composite scores (Table 4).

In addition, we investigated whether the composite scores predicted the accuracy of participants' first commitments. The analysis revealed no significant effect of the composite score in either of the experiments (Table 4).

[42] In Experiment 1, the majority of the participants (73) filled out the adaptive version of the BNT while 24 participants completed the standard four-question format of the BNT. The performance of the 'standard' group was calculated as if they had filled out the adaptive version of the BNT. The numeracy scores of the standard and adaptive group did not differ significantly ($M_{adaptive}$ = 2.23, $SD_{adaptive}$ = 1.12; $M_{standard}$ = 2.21, $SD_{standard}$ = 1.22), $t(36.84) = -0.09$, $p = .93$.

Table 4

Results of Generalized Linear Mixed Models Using the Cognitive Capacity Score to Predict the Correctness of the Initial and Final Response in the incongruent condition in Experiment 1 and 2

Predicted variable	Experiment	*OR*	95% CI	*z*	*p*
Correctness of final response	1	1.43	[1.13, 1.81]	3.01	.003
	2	1.58	[1.30, 1.93]	4.62	< .001
Correctness of initial response	1	1.05	[0.93, 1.18]	0.80	.424
	2	0.98	[0.90, 1.07]	-0.45	.656

Note. All models are generalized linear mixed models with a binomial link function.

Finally, we aimed to assess whether higher capacity individuals give more accurate final answers either because they make fewer changes when their initial answer is correct or because they are more likely to change their mind when their initial answer is incorrect – or both. Accordingly, we built a generalized linear mixed model testing whether the composite score predicted the number of CoMs when the initial answer was correct, and another model testing the relationship of composite score and CoMs when the initial answer was incorrect. As can be seen in Table 5, the analyses in both experiments revealed significant main effect of the composite score on the CoM: higher capacity people made more normative CoMs and less non-normative CoMs.

Table 5

Results of Generalized Linear Mixed Models Using the Cognitive Capacity Score to Predict the Occurrence of a Change of Mind Depending on the Correctness of the Initial Answer in the Incongruent Condition in Experiment 1 and 2

Initial answer	Experiment	*OR*	95% CI	*z*	*p*
Incorrect	1	1.45	[1.11, 1.90]	2.75	.006
	2	1.77	[1.42, 2.20]	5.06	< .001
Correct	1	0.67	[0.47, 0.95]	-2.28	.023
	2	0.70	[0.53, 0.93]	-2.43	.015

Note. We used generalized linear mixed models with a binomial link function.

Discussion

In two experiments, we aimed to explore the dynamics of people's decision-making to better understand why some individuals are more susceptible to biased thinking than others. We applied a novel mouse-tracking analysis technique to track people's first answer and thinking dynamics in reasoning situations without interrupting the reasoning process or relying on self-report measures. Using this method, we investigated the assumption that reasoners initially produce an incorrect answer in HB tasks. We observed that even in the incongruent trials of the denominator neglect task individuals move the mouse cursor first toward the correct response option in a substantial number of cases. This finding provides converging evidence with the result of recent studies using different methods (such as two response paradigms or thinking aloud protocols; see e.g., Bago & De Neys, 2017; Szaszi et al., 2017; Thompson and Johnson, 2014; Thompson et al., 2017) that people sometimes produce correct initial responses in HB tasks and that not everyone begins with a commitment to the incorrect response.

This finding suggests that models describing processes and individual differences in HB tasks need to explain and integrate the existence of the correct first responses. Bago and De Neys (2017) proposed that neither the classic default-interventionist dual process theory nor the classic parallel dual process models can account for this pattern and that such results are most aligned with a hybrid-model[43]. Their hybrid model suggests that several initial, intuitive responses (correct and/or incorrect) can be generated simultaneously, and their absolute strength will determine which of them will be used as the first answer. If the strength of the correct alternative is stronger, peoples' first answer will be correct (for an alternative hybrid model, see Pennycook, Fugelsang & Koehler, 2015). Note, however, that we cannot differentiate between the hybrid, default or parallel dual process models based on our results, since we cannot test one crucial element: which answer was generated by intuition and which by deliberation. Future research is needed to address this issue.

Travers et al. (2016) applied a similar mouse-tracking analysis to investigate the time-course of conflict in the CRT. In their paradigm, 4 different response options were presented to the participants and the authors analyzed the mouse trajectories to determine the sequence in which reasoners considered the response options. According to the model supported by their experiment, participants move the mouse-cursor towards the incorrect 'heuristic' option before

[43] For a detailed discussion on the comparison of dual-process models in light of a correct, intuitive first answer in HB tasks, see Bago and De Neys (2016).

choosing the correct option. Based on this result, the authors concluded that the CRT tasks automatically trigger a heuristic response which has to be suppressed in order to respond correctly. One might be tempted to infer that these results contradict our findings, but note that their results do not imply that reasoners never start with the correct response. Similarly, our findings do not indicate that the participants never had an incorrect first answer. The data suggests that although in the majority of incongruent trials individuals move the mouse cursor towards the incorrect response, sometimes they are first committed to the correct response.

We investigated another important aspect of individual differences in the HB tasks: the time point at which the capacity-normativity relationship arises. In contrast to Thompson and Johnson (2014), we did not find evidence for the idea that the high capacity reasoners produce more correct first answers[44]. Instead, we found that differences in performance between high and low capacity people arise after the first response is formulated. This finding is in line with the predictions of previous frameworks (De Neys & Bonnefon, 2013; Evans, 2007; Kahneman & Frederick, 2002; Stanovich & West, 2008). We observed that the deliberation after the first response benefited the higher capacity people in two ways: they changed their mind more often after an incorrect first response, and they changed their first response less often if it was correct. To our knowledge, this is the first empirical study showing that the latter effect also contributes to the capacity-normativity relationship[45].

Further research needs to investigate the exact role of previously identified causal mechanisms such as differences in the storage, monitoring ability, inhibition of the first response (De Neys & Bonnefon, 2013), feeling of conflict (Pennycook, Fugelsang & Koehler, 2015), qualitative versus quantitative differences in deliberation (Evans, 2007) or answer verification (Szollosi, Bago, Szaszi, & Aczel, 2017) which potentially drive the advantage of these late processes.

---

[44] The present study differs in several ways from Thompson and Johnson's (2014) study, which makes it difficult to identify what caused the discrepancy in the findings. First of all, we used a mouse-tracking measure to track peoples' first answers. Additionally, we employed different measures to estimate the participants' cognitive capacity, and used a version of the denominator neglect task which did not include the pictures of trays. Finally, we applied a different statistical analysis approach.

[45]However, it is worth highlighting that this effect is relatively small. In the present studies, participants made incorrect CoMs only in 3% of the trials in both study 1 and 2. Bago and De Neys (2016) found in 4 experiments that in 2%-6% of the trials participants changed from correct to the incorrect answer in the base rate and syllogisms tasks, while Szaszi et al. (2017) observed that the same value was 0.02% in the CRT tasks.

We think that the AOI mouse-tracking analysis technique can provide an additional way to test important questions in the reasoning literature and has some important advantage in tracking the first answer. First, in contrast to studies where participants are aware that the process of their thinking is tracked, mouse-tracking is much less obtrusive and might therefore decrease the likelihood that participants try to deliberately hide their dynamics of thinking. This is especially important if participants are not confident about their intuitions or strive to appear more competent. Secondly, the AOI mouse-tracking technique can assess the initial response without interrupting the decision process. Therefore, mouse-tracking based methods can be especially useful and sensitive tools to track individuals first commitments and choice tendencies.

A key underlying assumption of the employed mouse-tracking paradigm is that if a choice option (i.e., response) is activated in the reasoners mind, she will move the mouse towards the activated option[46]. However, a critic might argue that we cannot exclude unequivocally that the reasoners deliberatively suppressed an activated heuristic answer before they started to move their mouse cursor. Although we accept this possibility as a limitation of our findings, we argue that our paradigm appeared to be at least to some degree sensitive to track initial responses given that the reasoners were more likely to first move the mouse towards the incorrect response option in incongruent than in congruent trials.

Three more issues need to be considered in relation to our findings. First, similar to previous research investigating the capacity-normativity relationship, we cannot make conclusions on which component of cognitive capacity caused the observed effects in our study. Second, it also remains for future research to explore what effect cognitive style has on the dynamics of thinking in HB tasks. Finally, since previous studies suggested that HB tasks are not as homogenous as previously thought (Aczel, Bago, Szollosi, Foldes, & Lukacs, 2015), future studies should explore how the present findings generalize to other tasks or even to other versions of the denominator neglect task. Although the present results indicate that more accurate responding of higher capacity individuals in the denominator neglect task generally arises from either the override of the first response if it was incorrect or the less frequent change of the first response if it was correct, it is safe to hypothesize, that the model supported in this

[46] Or at least will be more likely to move the mouse towards the activated option than towards the non-activated one.

paper is not going to work everywhere[47]. In some contexts, some individuals give biased answers because they produce quick incorrect responses, while in other cases the biased answer is rather the results of a lack of deliberate thinking. Future research needs to create a taxonomy and determine the personality and task features which lead one or the other type of bias to dominate.

In the present research, we studied how individuals differ in their ability to provide normative responses and tested some of the key predictions of the models describing individual differences in HB tasks. Using a novel mouse-tracking analysis technique (the AOI), we consistently found that individuals produce both correct and incorrect first answers in the denominator neglect task. Furthermore, the capacity-normativity relationship seemed to arise late in the decision-making process in line with the predictions of several decision-making models; that is, we did not find evidence that higher capacity individuals had more correct initial answers but observed that reasoners corrected their first answer more often if it was incorrect. Moreover, we observed that high capacity individuals made fewer changes after correct first answers. Our study showcases how mouse-trajectory analysis can be utilized to investigate individual differences in decision-making and its results can help better apprehend the time-course of thinking and individual differences in HB tasks.

Data and Analysis Code

The analysis code and the underlying databases are openly available at http://www.sjdm.org/journal/vol13.1.html.

Acknowledgments

Barnabas Szaszi was supported by the Hungarian Fulbright Committee, the Rosztoczy Foundation and the Magyar Állami Eötvös Ösztöndíj (MÁEÖ2016_27). He is also grateful to Melinda Szászi-Szrenka for her supporting love and patience throughout the study. Bence Palfi is grateful to the Dr Mortimer and Theresa Sackler Foundation which supports the Sackler Centre for Consciousness Science. Aba Szollosi was supported by the "Nemzet Fiatal Tehetségeiért" Scholarship (NTP-NFTÖ-16-1184).

[47] See, for example, that the default-interventionist view wasn't supported in the in the domain of moral judgments (Gürçay & Baron, 2017; Koop, 2013).

Supplementary Materials

Assessing the initial and subsequent choice commitments using mouse-tracking

Previous research mostly applied mouse-tracking to simple problems (e.g., lexical decision making, social categorization) (implicitly) assuming that decision makers don't make more than two choice commitments in a given trial. If this assumption holds, the initial commitment and the number of choice commitments (one or two) may be identified using a cut-off threshold on the maximum deviation (MD) of the trajectories (see Freeman, 2014). If the MD value is higher than the threshold, it is assumed that there are two choice commitments in the trial and the initial commitment differs from the final answer (there is a change of mind, CoM). Meanwhile, if the MD value is lower than the threshold, it is assumed that there is only one choice commitment and there is no CoM in the trial. However, this approach does not generalize to trials with more than two choice commitments and can lead to biased conclusions regarding the individuals' first commitment if there are more than two commitments in a given trial (see Figure 1D). The AOI analysis technique was developed with the aim to overcome this limitation. We expected that similarly to Travers et al. (2016) in a considerable number of trials more than two choice commitments would be identified. In line with our expectations, our results showed that in a substantial number of trials in the ratio bias task people had more than two choice commitments (Table S1).

Table S1

Number of trials by number of choice commitments separately for each experiment

N of commitments	1	2	3	4	5	6
N of trials in Experiment 1	3038	1576	351	43	5	0
N of trials in Experiment 2	3580	2351	614	135	15	3

Is Action Execution Part of the Decision-Making Process? An Investigation of the Embodied Choice Hypothesis (Study 3)

Balazs Aczel[1], Aba Szollosi[1,2], Bence Palfi[1,3,4], Barnabas Szaszi[1,5], Pascal J. Kieslich[6]

[1]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[2]School of Psychology, University of New South Wales, Sydney, Australia
[3]School of Psychology, University of Sussex, Brighton, UK
[4]Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK
[5]Doctoral School of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[6]Experimental Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany

Abstract

In this study, we aimed to explore whether action execution is an inherent part of the decision-making process. According to the hypothesis of embodied choice, the decision-making process is bidirectional as action dynamics exert their backward influence on decision processes through changing the cost and value of the potential options. This influence takes place as moving towards one option increases the commitment to and, therefore, the likelihood of choosing that option. This commitment effect can be the result of either (a) the continuous act of getting closer to this option, or (b) the increased movement cost associated with changing the movement direction to select a different option. To disentangle the potential influence of these two factors, we developed the Guided Movement Task, a choice task designed to bias participant's computer-mouse movements by constraining the allowed movement space by a corridor. Using this task, we created different conditions in which the participants' mouse-cursor, after being guided towards one of the options, either had equal or unequal distances to the choice options. By this manipulation, we could test whether the continuous act of getting closer to an option in itself is sufficient to influence people's decisions – a claim of "strong embodiment". In two experiments, we found that the likelihood of choosing an option only increased when the distances between the two options were unequal after the initial movement, but not when they were equal. These results disagree with the hypothesis that action execution is an inherent part of the decision-making process.

Introduction

Imagine that you are looking for a quiet table in a busy café and you spot one on the far side. While walking towards the place and contemplating if you should spend your time there, you see a more peaceful-looking place on the other side. In the end, you choose the one you have been already moving towards. Did your initial movement towards the first option increase your preference for it or did the additional effort required to go to the second place make you stay with your otherwise less preferred option?

Most traditional decision-making models describe a decision as either a serial or a parallel process where action execution has no backward influence on the decision-making process. According to the serial models, decisions need to be completed before the initiation of their execution (Newell & Simon, 1972). Parallel models, in contrast, claim that there is a continuous flow from the decision processes to the action, and therefore, decision and action execution can occur in parallel (Coles, Gratton, Bashore, Eriksen, & Donchin, 1985). What both process models agree on is that the action is always the output of, and not the input to, the decision process.

It can be argued, however, that if the action requires some effort then it can influence people's choice. Laboratory experiments have provided clear examples of how choice commitment depends on the costs associated with choosing the available options. Resulaj, Kiani, Wolpert, and Shadlen (2009), for example, presented participants with a choice situation in a noisy visual environment in which they had to indicate their choices by moving a handle. The analysis of the choice trajectories showed an effect of cost: self-correction was dependent on the cost of modifying the trajectory; with increased movement cost, people adhered more to their present target. In a more direct test of this effect, the researchers varied the distance between the two options (Burk, Ingram, Franklin, Shadlen, & Wolpert, 2014). The results indicated that the frequency of changes of the initial choice is dependent on the excursion of the hand needed for moving to the opposite choice. Specifically, the frequency of changes decreased when the options were further separated. In sum, in these environments, movement towards a target changes the biomechanical costs of changing the target as it requires extra effort to alter the initiated movement towards a further option. These empirical findings highlight that a model of decision making cannot be ignorant of the movement costs associated with choosing between the available options.

Building on these observations, a recent hypothesis suggests that the decision-making process is not unidirectional, but rather that it is dependent on the action and its dynamics. In

their embodied choice model[48], Lepora and Pezzulo (2015) argue that in ecologically valid scenarios decisions and actions are in a continuous loop since our actions can influence those sensory stimuli that shape our decisions. For example, in the café, you might already be approaching your first option before you have been fully committed to choosing it, but your movement towards that option in itself increases your commitment to the target. Supporting empirical evidence for this model would mean that action dynamics are not just results, but also inherent parts of the decision process. This finding would also make the model an extension of the embodied theory of mind to the domain of decision making, increasing the importance of understanding and assessing the model predictions.

The model proposes that action dynamics exert their backward influence on decision processes through changing the cost and value of the potential options. Lepora and Pezzulo (2015) also provided a formalized description of how action dynamics influence the decision processes. In addition to the accumulated information, their model includes two novel characteristics, the current location and the current movement. The model suggests that the commitment effect will increase as the agents are closer to one of the targets and as they are "sufficiently engaged in moving towards a target" (p. 10). The authors argue that mouse tracking studies can provide ideal data for testing this hypothesis. In their study, they analyzed whether serial, parallel, or embodied choice models give a better account of previous data from mouse tracking experiments. In these assessments, the model fitted best the empirical observations when it took into account a commitment parameter reflecting how the choices became more biased as the mouse moved closer to one of the targets.

In the present paper, we argue that studying movements in unconstrained space does not allow us to differentiate whether moving towards one choice option increases the commitment effect as the result of either the continuous act of getting closer to this option, or the increased movement cost[49] associated with changing the movement direction to select a different option. Movement in unconstrained spaces, such as moving the cursor on the computer screen, leaves these two factors confounded. In such designs, it cannot be established that it is the increased cost of the movement alone that is influential on decision making (which is needed to move

[48] The definition of embodied cognition is an unsettled question of the literature. For a detailed discussion of its relevant aspects see the *General Discussion*.

[49] Note that unequal distances to the options can result not just in imbalance in movement costs but also in differences in the time needed to reach either option. Although the model emphasizes the effect of the biomechanical cost, the longer movement time might also discourage the agent to approach the more distant target.

towards the alternative, ultimately not selected option), and not the influence of bodily movement as well. If this is true, then it could be possible that the movement itself does not influence the decision process, but that only the difference in distances and consequently the difference in movement costs associated with choosing either option influence people's willingness to execute the decision. In this case, the general *law of least effort* (Hull, 1943) could be a viable alternative to explain the effect, as humans (and other animals) seek to spend as little physical or mental effort as possible and, therefore, prefer to choose the option associated with the least movement cost (Zipf, 1950).

To disentangle the potential influence of the two factors, an experimental design is required which allows the agent to execute a continuous movement towards one of the options after which the distance to the options, and as a consequence the movement cost associated with choosing either option is equal. In the present study, we introduce the Guided Movement Task (GMT), which biases the way participants move a computer mouse cursor through constraining the allowed movement space by a virtual corridor. In doing so, the GMT allows the investigator to influence the direction of the participants' initial movement while controlling the distance of the cursor from the choice options. Importantly, these manipulations can create a situation in which the initial movement is directed towards one of the choice options, but the distances from each of the choice options (and hence the movement costs associated with choosing them) are the same at the time in which participants indicate their decision (see Figure 1C and D below). In this situation, the participant's choice can only be influenced by the direction of the initial movement, but not by the difference in movement costs.

The aim of the present study was to test the embodied choice hypothesis in mouse movement environments. Specifically, we tested whether movements towards the options in themselves have an influence on peoples' choices. To do this, we used the GMT design with choice situations in which the first movements of the mouse-cursor were directed towards one of the options after which the distance to the options was either equal or unequal. In both experiments, we used preferential choice tasks because choices on such tasks are not verifiable and as such, they are possibly more prone to the subtle influence of the movement compared to when a choice is more straightforward. We hypothesized that if the direction of the initial movement itself influences the participant's choice then this influence should be present even when the distances to the two options and, consequently, the movement costs, are kept equal by the design.

Experiment 1

Methods

Participants

We recruited 115 university students[50] (98 females; *M* = 22.36 years, *SD* = 2.04 years) through a course at Eotvos Lorand University (Budapest, Hungary), all of whom were native Hungarian speakers. The participants received course credit in exchange for participation. The research was approved by the institutional ethics committee of Eotvos Lorand University, Hungary.

Guided Movement Task

The current study employed a new task, the GMT, in which participants had to initially follow a predetermined movement corridor with their mouse cursor before indicating their choice by clicking on one of two options. This movement corridor was varied between trials in order to manipulate the distance to the options after the initial movement and, in addition, disentangle this from the initial movement direction. In a within-subjects design, two factors were manipulated through the corridors. First, the *direction* of the initial movement was biased either towards the right (Figure 1A and 1C) or towards the left (Figure 1B and 1D) option. Second, the *distance* to the options after the initial movement was manipulated by placing the end of the corridor in the middle of the screen (equal distance), or in one of the corners beneath one of the options (unequal distance). Figure 1A and 1B show the conditions in which the distance to the options is unequal at the end of the movement corridor since in 1A the distance to the right option and in 1B the distance to the left option is reduced. In contrast, Figure 1C and 1D reflect the conditions with equal distance to either option at the end of the movement corridor. Consequently, in the latter two conditions only the direction of the initial movement could bias the choice of the options, whereas in the first two conditions additionally the movement cost associated with choosing either option differs.

[50] Two additional participants had to be excluded due to problems with their computer screens during the experiment.

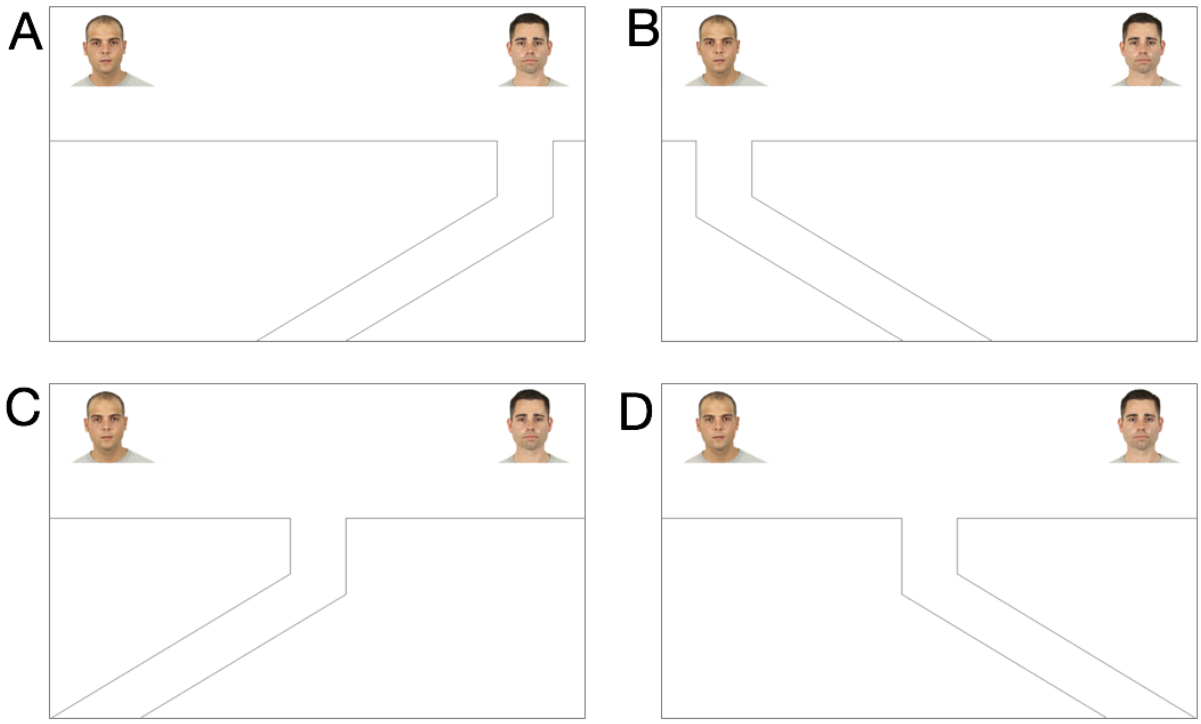

Figure 1. Summary of the experimental manipulations in Experiment 1. The distances to the options after the initial movement are unequal in panel A (shorter distance to the right option) and B (shorter distance to the left option), and equal in panel C and D. The initial movement is directed towards the right option in panel A and C, and directed towards the left option in panel B and D.

Stimuli

For the first experiment, 180 portrayal pictures (90 male, 90 female) were gathered from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015) with the permission of the authors. In each picture, the person wore a gray sweater and was photographed in front of a white background. The pictures were used in pairs in the GMT and participants were instructed to choose the picture that they found more attractive.

To enhance the probability that potentially biasing effects of the movement cost and initial movement direction would be revealed, we made the choices more difficult by pairing the pictures of similarly attractive faces. Attractiveness was assessed in a pilot study, where 102 participants rated each portrait on a 10-point Likert scale. Female and male portraits were separately ordered according to their average attractiveness ratings and the adjacent portraits were grouped into the choice pairs for the experiment. The faces were paired by gender in order to increase the likelihood that participants compare the pictures along similar dimensions.

Procedure

The experiment was implemented in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012) and run in full-screen mode with a computer screen resolution of 1440x900 pixels. Mouse movements were recorded using the mousetrap plugin for OpenSesame (Kieslich & Henninger, in press). Default mouse-sensitivity settings of Microsoft Windows 10 were used. After providing informed consent and demographic information, the participants were instructed that they would see pairs of faces and their task was to choose the more attractive one by clicking on it with the computer-mouse ("Which one of the following faces do you find more attractive?"). Additionally, participants were informed that in order to make the experiment more interesting they would have a secondary task. The task was not to cross or touch the sides of the corridors (displayed in Figure 1) during the experiment. To incentivize this behavior, additional course credit was offered to those who made errors in less than 10 percent of the trials.

First, the participants practiced the task on four trials (two pairs of pictures were employed from the collection in this section of the experiment in which each pair was presented twice and not used in the actual task). After each practice trial, participants received feedback informing them whether or not they had left the prescribed corridor with the cursor. After the practice trials, the participants were presented with 88 experimental trials in random order and without feedback (88 pairs of pictures were used, each pair being only presented once). For each participant, each of the 88 pairs was randomly assigned to one of the four experimental conditions (outlined in Figure 1), resulting in 22 trials per condition. Additionally, the position of the pictures on the screen (left/right) was randomized for each trial. Participants' choices, response times, and the coordinates of their mouse movements were continuously recorded. The mouse movements were utilized to assess whether the cursor had left the prescribed corridor in a trial. In the second part of the experiment, participants were asked to rate the attractiveness of each picture on a 10-point Likert scale (these data were not analyzed in the current study). At the end of the experiment, participants were asked to answer the following question: "Please write down if you had any thoughts during the experiment about what we aimed to study in this research".

Hypotheses

As both the stimuli and the experimental conditions were fully counterbalanced between the participants, systematic difference in participants' choices between conditions can only be caused by the experimental manipulations. Specifically, we expected that if only distance inequality has an influence on participants' choices, the proportion of choices of the right-side option should be higher if the corridor ends beneath the right option and lower if it ended beneath the left option, while there should be no systematic differences between the two conditions with balanced movement costs. However, if the direction of the initial movement itself influences participants' choices, there should also be an influence of movement direction in the equal distance conditions with a higher proportion of choices of the right-side option if the initial movement is heading towards the right option and a lower proportion of such choices if it is heading towards the left option.

Results

As a first step in the analysis, we excluded the trials in which participants left the corridors (7.2% of trials). To get a descriptive overview of the data, we calculated the percentages of choices of the right-side option per participant for each *direction* condition (initial movement towards the left vs. right option) separately for each *distance* condition (distance to options equal vs. unequal after the initial movement). In the conditions with unequal distances, the mean percentage of right-side option choices was higher in the direction right condition (61.9%) than in the direction left condition (42.7%). In the conditions with equal distances, the mean percentages of right side option choices were comparable for the direction right (53.6%) and the direction left (55.6%) conditions.

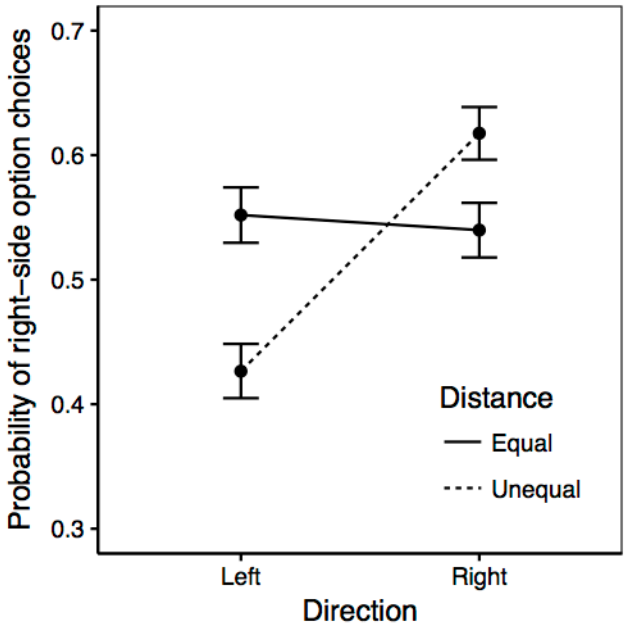

Figure 2. The probability of right-option choices in Experiment 1 depending on the initial movement direction (towards the left vs. right option) and the distances to the options after the intial movement (equal vs. unequal). Error bars represent 95 % confidence intervals based on a generalized linear-mixed model.

To statistically test this pattern, we built a generalized linear mixed model in which we predicted the probability of choosing the option on the right side with a binomial link function specifying a random intercept per participant. As predictors, direction (right vs. left), distance (equal vs. unequal), and their interaction were included. The direction and distance predictors were dummy coded and the condition with an initial movement direction towards the left option and equal distances after the initial movement served as the baseline. In the analyses, the reported Odds Ratios (*OR*) represent the relative change of the odds of choosing the option on the right side in the coded condition compared to the specified baselines. For the analyses we utilized the glmer function of the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015). Table 1 displays the results of the model.

Table 1

Results of the Generalized Linear Mixed Model Predicting the Probability of Choosing the Option on the Right Side in Experiment 1

Predictors	***OR***	**95% *CI*s**	***z***	***p***
Intercept	1.23	[1.13, 1.35]	4.55	< .001
Effect of direction (right vs. left, in equal distance condition)	0.95	[0.85, 1.07]	-0.82	.41
Effect of distance (unequal vs. equal, in direction left condition)	0.60	[0.54, 0.68]	-8.46	< .001
Interaction of direction and distance	2.28	[1.93, 2.69]	9.82	< .001

Note. A binomial link function was used and a random intercept per participant was specified. Predictors were dummy coded (with direction left, equal distance as baseline condition).

As Table 1 and Figure 2 show, the effect of direction was not significant in the baseline condition when the distances to the options were equal, $OR = 0.95$, 95% CI [0.85, 1.07], $z = -0.82$, $p = .41$. Furthermore, the effect of direction was significantly larger when the distances were not equal, as indicated by the interaction, $OR = 2.28$, 95% CI [1.93, 2.69], $z = 9.82$, $p < .001$. Accordingly, the effect of direction was significant in an additional model where the unequal distance, direction left condition served as the baseline, $OR = 2.17$, 95% CI [1.93, 2.44], $z = 13.03$, $p < .001$.

In addition, we computed Bayes factors (*B*) using the lmBF function of the BayesFactor package (Morey, Rouder, & Jamil, 2015) with default priors modelling the predictions of the alternative hypotheses. A Bayes factor is a measure of relative evidence and it can quantify the support that the collected data provide for a certain hypothesis over another hypothesis (e.g., Dienes, 2014). Thus, one of the advantages of the Bayes factor analysis over the null-hypothesis-significance-testing approach is that it differentiates between the cases of data-insensitivity and null results (Morey, Romeijn, & Rouder, 2016). We will report *B* as the ratio of the likelihood of the data under H1 compared to under H0, and so a *B* value higher than 1 can be interpreted as evidence favoring the alternative hypothesis whereas a *B* value lower than 1 can be interpreted as support for the null. To compute Bayes factors for the direction effect in

each of the distance conditions we aggregated the choice rates for each condition per participant and performed a linear model for each distance condition using direction as predictor and treating participants as a random factor. The analyses indicated support for the effect of direction in the unequal distance conditions, $B = 1.81 * 10^{15}$, and provided evidence for the absence of this effect in the equal distance conditions, $B = 0.30$.

We further examined whether the initial movement direction affected response latencies. Although it is possible that the influence of the movement towards the options was not strong enough to influence participants' choices in the equal distance conditions, conflicting action plans could have still modified the speed of the execution. In doing so, we calculated choice-bias congruency such that the trials when people chose the option that was in the direction of the initial movement were coded as congruent, and trials when people chose in opposition to the direction of the initial movement were coded as incongruent. We then tested separately for the equal distance and the unequal distance conditions if congruency influenced the response times (RT) in a linear-mixed model including random intercepts for each participant. To do this, we used the mixed function from the afex package (Singmann, Bolker, & Westfall, 2015). In the unequal distance conditions, congruent choices led to significantly shorter RTs, $b = -627$, 95% CI[-708, -546], $t(4646.72) = -15.22$, $p < .001$, reflecting the fact that the distances to the options in these conditions were indeed unequal at the end of the movement corridor as participants took longer to move toward the more distant option. Crucially, in the equal distance conditions, choice congruency did not significantly affect RTs, $b = -40$, 95% CI [-126, 46], $t(4534.41) = -0.91$, $p = .36$. Bayes factors using the per participant and condition aggregated RTs indicated support for the effect of congruency in the unequal distance conditions, $B = 1.54 * 10^{16}$, and evidence for the absence of this effect in the equal distance conditions, $B = 0.23$.

A potential explanation for the absence of the effect of congruency in the equal distance conditions could be that participants used more time to plan their decision in incongruent than in congruent trials before beginning to move the mouse. To test this idea, we computed the movement initiation times using the mousetrap R package (Kieslich, Wulff, Henninger, Haslbeck, & Schulte-Mecklenbeck, 2016) and compared them for congruent and incongruent trials in the equal distance conditions. Again, we found evidence for the absence of an effect of congruency (see Supplementary Materials for the details of this analysis).

To examine the robustness of the results, we repeated the analyses excluding participants whose reports about the hypothesis included a reference to the corridors or to the

mouse movements (14 of 115 participants). The exclusion of these participants did not change the pattern of results of any analyses. Furthermore, the response time analyses could also be replicated with their log-transformed variant (see the Supplementary Materials for more details).

Discussion

In this experiment, we applied a new type of task, the GMT, to investigate whether moving towards an option increases the likelihood of choosing that option. The results indicated that the direction of the initial movement alone did not significantly affect participants' choices and the corresponding Bayes factor indicated support for an absence of the effect. Participants' choice patterns significantly differed between conditions only when the initial movement direction led to unequal distances to the options at the end of the movement period.

It could be possible that while moving towards an option can activate action preparation, this effect is not strong enough to modify peoples' choices. In this case, we would still expect that the competing action plans slow down action execution in cases where a participant ultimately decides against the option the initial movement was directed towards. However, our response time analysis found no support for this possibility; in fact, the data supported the null-hypothesis. These results suggest that among the two potential effects, only a difference in distance to the options, and not the direction of the initial movement itself, influences people's choices.

Experiment 2

To be able to generalize our findings, in Experiment 2, we replicated the present design using a non-visual preference task. Risky choice tasks require people to decide between different options that lead to outcomes that occur with certain probabilities. In many studies, participants decide between pairs of monetary lotteries whereby one lottery involves a large and risky reward, and the other a smaller and less risky reward. The task was ideal for our purposes, as people's risk-taking preference lies on a continuum and can potentially be the subject of weaker-level influences (e.g., Arkes, Herren, & Isen, 1988; Gardner & Steinberg, 2005). We expected that if the movement itself can bias the decision process, then people would display an increased preference towards those options to which they were moving towards

initially during the trial. In other words, the direction of the virtual corridor should have an effect on people's choice, even when we keep the movement demand equal for reaching the options.

Methods

Participants

We recruited 194 students (150 females; M = 21.82 years, SD = 1.80 years) from the local subject pool of Eotvos Lorand University. All participants were native Hungarian speakers and provided informed consent before the study.

Stimuli

The stimuli consisted of 43 gamble pairs created by Koop and Johnson (2013, Study 2). Each gamble entailed an outcome value and a probability estimate of success (e.g., win $70 with 80%, otherwise nothing). The pairs were created in such a way that while one gamble had a higher outcome value and the other had a higher success probability, both options had an approximately equal expected value.

Design and Procedure

The design and the procedure followed that of Experiment 1 except for the changed material. As displayed in Figure 3, in Experiment 2, the participants' task was to indicate which of two monetary gambles they would choose ("Of the following two options, which one would you prefer to choose?"). Out of the 43 gamble pairs, 40 were presented during the actual experiment, one gamble pair was shown during the instruction section, and the gambles from the remaining two pairs were used in the practice part twice. The 40 trials (10 trials per condition) were presented to each participant in the experimental part in random order and the position of the gambles on the screen (left/right) was randomized for each trial.

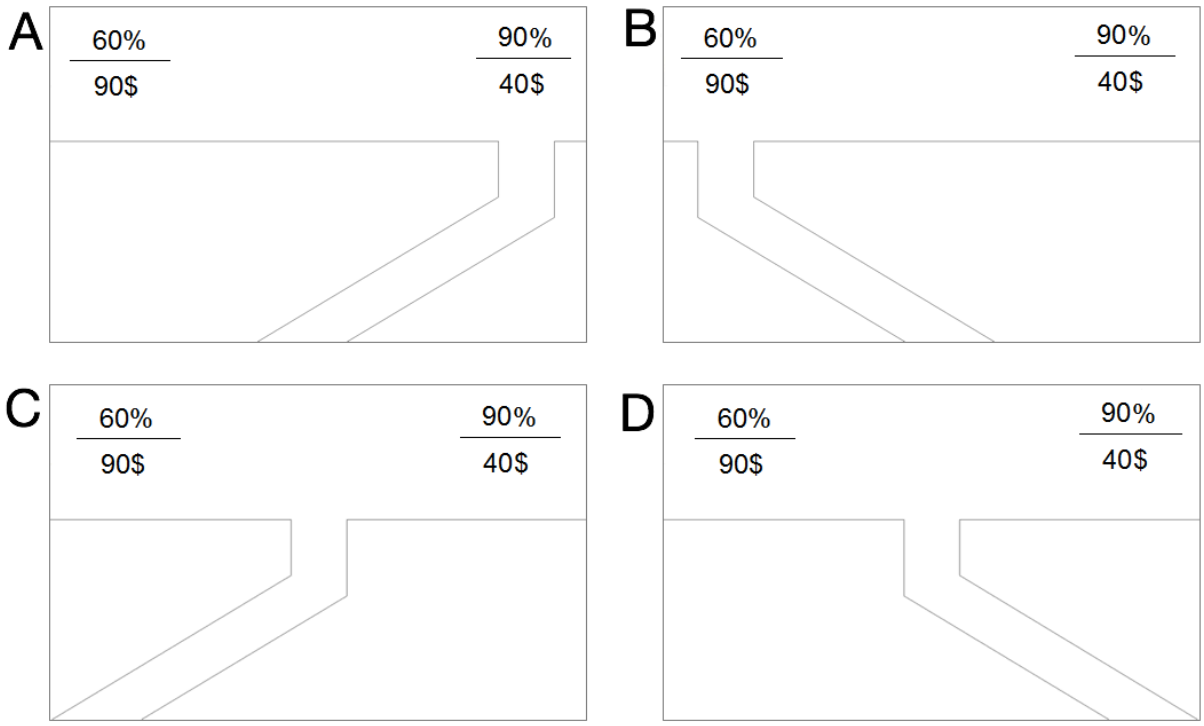

Figure 3. Summary of the experimental manipulations in Experiment 2. The distances to the options after the initial movement are unequal in panel A (shorter distance to the right option) and B (shorter distance to the left option), and equal in panel C and D. The initial movement is directed towards the right option in panel A and C, and directed towards the left option in panel B and D.

Results

The analyses in Experiment 2 closely followed those of Experiment 1. Firstly, we excluded the trials in which participants left the corridors (4.7% of trials). We again calculated the percentages of choices of the right-side option per participant for each combination of the *direction* and *distance* conditions. The pattern was similar to that of Experiment 1, as in the unequal distance conditions the mean percentage of right side option choices was higher in the direction right condition (53.3%) than in the direction left condition (44.1%). In the equal distance conditions, the mean percentages of right side option choices were again comparable for the direction right (51.4%) and the direction left (53.0%) conditions.

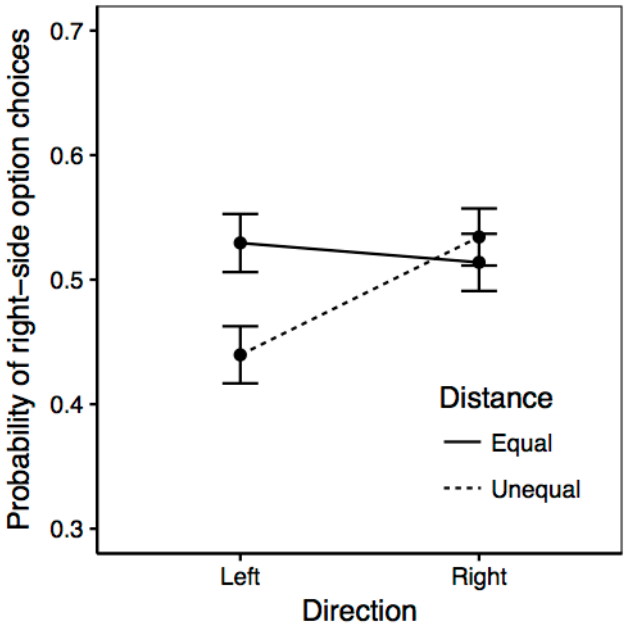

Figure 4. The probability of right-option choices in Experiment 2 depending on the initial movement direction (towards the left vs. right option) and the distances to the after the intial movement (equal vs. unequal). Error bars represent 95 % confidence intervals based on a generalized linear-mixed model.

To statistically test this pattern, we predicted the choice of the right-side option in a generalized linear model with binomial link function using initial movement direction (right vs. left), distance to the options after the initial movement (equal vs. unequal), and their interaction as predictors (and including a random intercept per participant). Predictors were again dummy coded with the equal distance, direction left condition as the baseline. The results of the model are summarized in Table 2.

Table 2

Results of the Generalized Linear Mixed Model Predicting the Probability of Choosing the Option on the Right Side in Experiment 2

Predictors	***OR***	**95% *CI*s**	***z***	***p***
Intercept	1.13	[1.02, 1.24]	2.47	.01
Effect of direction (right vs. left, in equal distance condition)	0.94	[0.83, 1.07]	-0.94	.35
Effect of distance (unequal vs. equal, in direction left condition)	0.70	[0.61, 0.79]	-5.42	< .001
Interaction of direction and distance	1.56	[1.30, 1.87]	4.74	< .001

Note. A binomial link function was used and a random intercept per participant was specified. Predictors were dummy coded (with direction left, equal distance as baseline condition).

As can be seen in Figure 4 and replicating the findings of Experiment 1, we found no significant effect of direction on the proportion of right-side choice in the equal distance conditions, OR = 0.94, 95% CI [0.83, 1.07], z = -0.94, p = .35. The significant interaction of distance and direction indicated that the effect of direction was significantly larger in the unequal distance conditions, OR = 1.56, 95% CI [1.30, 1.87], z = 4.74, p < .001. As in Experiment 1, the effect of direction was significant when using the unequal distance, direction left condition as the baseline, OR = 1.46, 95% CI [1.29, 1.66], z = 5.76, p < .001. Bayes-factor analyses using the aggregated choice rates per condition and participant indicated support for the effect of direction in the unequal distance conditions, $B = 1.84 * 10^4$, and evidence for the absence of the effect of direction in the equal distance conditions, B = 0.18.

Similar to Experiment 1, we analyzed the response latencies depending on participants' choice congruency. In the unequal distance conditions, RTs were again significantly lower in congruent compared to incongruent trials, b = -540, 95% CI [-662, -419], t(3562.05) = -8.70, p < .001, the Bayes factor supporting this effect, $B = 4.84 * 10^{11}$. Unlike in Experiment 1, however, RTs in the equal distance conditions were significantly higher in congruent than in incongruent trials, b = 196, 95% CI [85, 308], t(3501.00) = 3.45, p < .001, the Bayes factor indicating support for the alternative, B = 423. Importantly, however, the effect was in the opposite direction of what one would have predicted when assuming a biasing influence of the

initial movement direction. Congruency again had no effect on the movement initiation time in the equal distance conditions (see Supplementary Materials).

The results could be replicated when excluding the participants whose reports mentioned the corridors or the bias of the mouse movements (22 of 194 participants). Repeating the analysis on the log-transformed response times also replicated the original patterns (see the Supplementary Materials for more details).

Discussion

In this experiment, we repeated the design of Experiment 1 with a non-visual preference task. The motivation for this experiment was to test whether the pattern of results that we found on the face-preference task could be replicated in a different kind of choice task. Again, we found that the participants' choices only differed significantly between conditions when the distances to the options were unequal after the initial movement. When the distances were equal, no significant effect of direction was observed. Furthermore, the Bayes factor provided evidence for the absence of the effect. Unlike in Experiment 1, the analysis of response time data revealed an effect of direction in the equal distance conditions; however, this effect was in the opposite direction compared to what the embodied choice model would predict, as responses were significantly faster in incongruent than in congruent trials. The pattern of results did not differ when we excluded those participants from the analysis who reported some insight about the purpose of the experimental design.

General Discussion

This research was motivated by the question of whether action execution is part of the decision-making process. More specifically, we aimed to explore whether the movement towards an option in itself has an influence on peoples' choices. To address this, we created a new experimental task, the Guided Movement Task (GMT), which allowed us to manipulate participants' initial movement direction while keeping the distances to, and as a consequence the movement costs associated with choosing either option equal. If the movement towards an option creates a commitment effect in itself, then we would expect an increase in participants' choices of the direction biased option even when the movement costs associated with choosing either option are equal. We tested this question in two experiments using a face-preference and a risky choice task. Across both experiments, the initial movement towards an option did not

bias participants' choices when the distances to the options were equal. This bias only occurred when the initial movement also led to an increase in distance to the other option. In other words, these results suggest that moving towards an option in itself, does not increase commitment, and it hence does not increase the preference of the option.

Previous behavioral studies showed that people are less likely to change their mind in the later part of the test trials (Kiani, Cueva, Reppas, & Newsome, 2014; Resulaj et al., 2009) and that the initial decisions decrease the sensitivity to subsequent information (Bronfman et al., 2015). It can be argued that in real-life situations this decision commitment has adaptive functions. Ambiguous evidence can lead to indecisiveness and, in certain situations, time wasting can be more costly than holding onto the first choice, especially when the action changes present significant cost. From this aspect, the proposal of the embodied choice model is in accord with previous expectations. The question, however, is whether, in cases when bodily movement is part of the choice task, an increased commitment to a target is the result of (a) the movement towards the target and/or (b) the decreased distance from the target. Our results argue against the first interpretation and support the second possibility. It is important to highlight, however, that the effect of distance asymmetry is not necessarily the manifestation of the embodiment. While it is possible that the movement cost associated with the distance itself has a direct effect on the decision process by modifying the perceived values of the options, it is also plausible to suggest that people's preference towards the options does not change as a result of the associated movement costs, instead it only influences the willingness to act in accordance with the preference. In other words, it is possible that people move to the initially approached option not because they prefer it over the other option, but only to save the energy needed to move to the alternative option. This might especially be true for tasks of little personal importance, which arguably might hold for the studies we reported.

An unavoidable difficulty of any test of the embodiment phenomena is that the use and definition of embodied cognition vary across publications, covering a range from weaker to stronger claims. In this paper, we followed the definition used in the description of the embodied choice model (Lepora & Pezzulo, 2015) meaning that "action and its dynamics are considered an integral part of the decision-making process, rather than merely a way to report an already made choice" (p. 5). Strict adherence to this definition implies that embodiment is concerned only in case of some kind of bodily action. Biomechanical cost, in itself, is not an action, but the consequence of the action, therefore its influence cannot be taken as evidence

for embodied choice. To examine the effect of movement in itself, action and its biomechanical cost have to be separated which we aimed to do in the present study.

Nevertheless, compared to this stronger claim, weaker interpretations of embodied cognition could provide the basis for an argument that the effect of movement cost is already evidence for the phenomenon. This would be the case, for example, if we assume that cognition is embodied if it is *"influenced by the body and the brain interacting with the environment"* (Glenberg, Witt, & Metcalfe, 2013, p. 573). As properties and affordances of the surrounding environment can influence choice behavior in ample ways, the effect of movement cost, or simply being closer to one of the options could be used for this argument. The question, however, remains as to how much we gain by introducing embodiment into the explanation of the behavioral phenomenon. Critics argue that most claims of weak embodied cognition can be equally accommodated by purely cognitive accounts (Mahon, 2015). As such, it can be contemplated whether the well-known sunk cost effect (Arkes & Blumer, 1985) could suggest a cognitive account for why people would choose the target they have been moving towards. Similarly, it could be argued that the biomechanical cost of moving to the option is simply a factor of the cognitive operation of choice, not qualitatively different from other factors such as attractiveness or estimated risk. Therefore, not the bodily actions, but their value in the cognitive model would affect the decision-making process.

The idea of embodied cognition has received fierce criticism in the last years from both logical and empirical perspectives (Durgin et al., 2009; Firestone, 2013; Goldinger, Papesh, Barnhart, Hansen, & Hout, 2016; Mahon, 2015; Mahon & Caramazza, 2008; Papesh, 2015). Nevertheless, we believe that there are viable research questions concerning the relationship of bodily movements and choice behavior. Decision making is a key aspect of our interaction with the environment and therefore, it has central importance to understand whether action should be incorporated into our models of decision making. The GMT design, unlike previous choice or mouse-tracking studies, makes it possible for researchers to systematically explore the relationship of mouse movement direction and movement cost on preference in two-option choice tasks.

While the GMT design offers a first opportunity to test and disentangle the specific effects of movement direction and movement cost, the design also has some potential limitations. First, the GMT is a computerized task in which participants use the mouse to follow a virtual path. While this is in line with previous tests of the embodied choice model that also relied on mouse movements (Lepora & Pezzulo, 2015), it would be useful to supplement the

current findings with additional studies that involve more direct manipulations of bodily movements, such as hand movements.

Additionally, one potential problem of the GMT is that the task design might induce demand effects. Specifically, a corridor pointing towards an option might be interpreted by the participants as a demand by the experimenter to choose that option. To address this concern, at the end of each study, we assessed the hypotheses that participants had about the purpose of the study. Importantly, the results of each study could be replicated after excluding participants that mentioned movement corridors or mouse movement biases. Nevertheless, we cannot completely exclude the possibility that differences in participants' choices in the unequal distance conditions were the result of such demand effects.

In addition, the corridors in the GMT ended with a short straight vertical section to prevent overshooting of the participants' movements in the direction of the biased option. It is possible that in this section of the corridor, a participant's movement could overshoot a bit towards the corridor boundary and that the participant corrects this afterwards by moving in the opposite direction. This might constitute a problem as it could weaken the direction effect. However, a potential reversal of the movement in the straight section of the corridor would only involve a fraction of the distance compared to the overall distance the participant initially travelled towards the biased option. Moreover, the potential reversals could occur both in the equal and unequal distance conditions, and, consequently, it cannot explain the observed differences of the direction effect between conditions.

In conclusion, by the help of a new testing design, the present investigation separately studied the effect of movement towards an option, and the effect of the cost of action execution. The fact that we found a biasing effect on choices only when distance to the options differed after the initial movement suggests that previous observations of commitment effects might have not been diagnostic to the strong embodiment hypothesis of choice. It remains a question of further research as to whether cost imbalance exerts its effect by changing the subjective values of the options (i.e., influencing the decision process), or by influencing people's willingness to initiate the execution of reaching the preferred target

Data and Analysis Code

The analysis code and the underlying databases are openly available at https://osf.io/etk8f/.

Acknowledgements

We would like to thank Garston Liang, Adrien Walker, and Melissa Wood for their comments on an earlier version of this paper, as well as Anna Palinkas, Marton Kovacs and Peter Szecsi for their assistance in data collection. Aba Szollosi was supported by the "Nemzet Fiatal Tehetségeiért" Scholarship (NTP-NFTÖ-16-1184). Bence Palfi is grateful to the Dr. Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science. Barnabas Szaszi was supported by the Hungarian Fulbright Committee and the Rosztoczy Foundation and the Magyar Állami Eötvös Ösztöndíj (MÁEÖ2016_27).

A Systematic Scoping Review of the Choice Architecture Movement: Towards Understanding When and Why Nudges Work (Study 4)

Szaszi, B.[1,2], Palinkas, A. [2], Palfi, B.[3,4], Szollosi, A. [5], Aczel, B. [2]

[1]Doctoral School of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[2]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[3]School of Psychology, University of Sussex, Brighton, UK
[4]Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK.
[5]School of Psychology, University of New South Wales, Sydney, Australia

Abstract

In this paper, we provide a domain-general scoping review of the nudge movement by reviewing 422 choice architecture interventions in 156 empirical studies. We report the distribution of the studies across countries, years, domains, subdomains of applicability, intervention types, and the moderators associated with each intervention category to review the current state of the nudge movement. Furthermore, we highlight certain characteristics of the studies and experimental and reporting practices which can hinder the accumulation of evidence in the field. Specifically, we found that 74 % of the studies were mainly motivated to assess the effectiveness of the interventions in one specific setting, while only 24% of the studies focused on the exploration of moderators or underlying processes. We also observed that only 7% of the studies applied power analysis, 2% used guidelines aiming to improve the quality of reporting, no study in our database was preregistered, and the used intervention nomenclatures were non-exhaustive and often have overlapping categories. Building on our current observations and proposed solutions from other fields, we provide directly applicable recommendations for future research to support the evidence accumulation on why and when nudges work.

Introduction

In the endeavor of helping people make beneficial decisions for society and for themselves, the implementation of *choice architecture* interventions became a promising approach. This approach was popularized by *Nudge*, the influential book of Thaler and Sunstein (2008) by orienting recent empirical research towards exploring how known deviations from rational behavior can be used to design choice situations that nudge decision-makers towards more beneficial options. The so-called '*Nudge Theory'* propagates the use of choice architecture techniques which are generally inexpensive and less invasive solutions compared to traditional, direct interventions, and build on people's habits and cognitive boundaries (Thaler, Sunstein, & Balz, 2014). From a policy-making perspective, this approach aims to improve the public good without the use of regulatory processes or economic incentives.

The popularity of the approach has resulted in a movement with an increasing number of research projects dedicated to explore the potential benefits of nudge techniques (Figure 1). Researchers tested the effectiveness of various modifications in the choice environments in several domains, such as transport (e.g., Gaker, Zheng, & Walker, 2010), financial decision making (e.g., Evans, Dillon, Goldin, & Krueger, 2011), public health (Li & Chapman, 2013; Vlaev, King, Dolan & Darzi, 2016), and sustainability (e.g., Baca-Motes, Brown, Gneezy, Keenan, & Nelson, 2013). For example, in a research project on sustainable energy use, 40,000 German households were asked whether they would purchase their energy from renewable sources for a slightly higher price or not (Ebeling & Lotz, 2015). Half of the households had to actively "opt-in" to make their decisions, while the other half had to "opt-out" in an online questionnaire. The results showed that while only 7% of households purchased green energy in the opt-in design, about 70% accepted the deal in the opt-out version. As choice architecture interventions are generally inexpensive and easily implementable, they received substantial interest in governmental policy making as well (e.g., Behavioural Insights Ltd, 2015; Dolan, Hallsworth, Halpern, King, & Vlaev, 2012; Social and Behavioral Science Team, 2015; Lourenço, Ciriolo, Almeida, & Troussard, 2016).

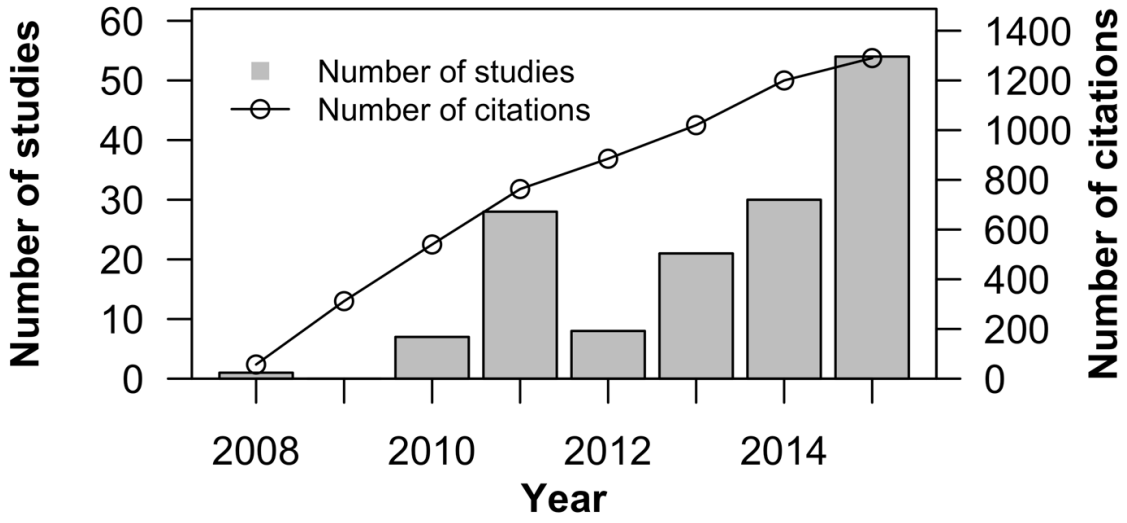

Figure 1. Number of studies included in our review and the number of citations on Thaler and Sunstein (2008) based on the search results in Google Scholar.

Choice architecture techniques seem to promise generally applicable and easily implementable solutions to important societal problems. To assess how well the accumulated evidence supports this impression, it is crucial to consult general and question-specific evaluations of the published studies. Systematic reviews of the scientific evidence concerning the specific nudging strategies are scarce. A few attempts have been made to give a systematic review of the effectiveness of choice architecture interventions within certain domain-specific areas (Bucher et al., 2016; Nørnberg et al., 2015; Skov et al., 2013; Wilson et al., 2016). From one aspect, these reviews revealed that the idea of nudging has inspired a large magnitude of studies for various applied questions. On the other hand, they also concluded that intervention designs and outcome measures are heterogeneous, and that the number of good quality empirical studies is low, which made it impossible to conduct meta-analyses on the effectiveness of choice architecture methods.

Skov and colleagues (2013), for example, provided a review with the focus on choice architecture aiming to change eating behavior in a self-service restaurant setting. As a result of a selective literature search of a 10 months long publication period, the researchers found 12 studies satisfying their inclusion criteria. The majority of these studies were found to have very weak quality due to small sample sizes, lack of descriptive information and inaccurate outcome measurement tools. Besides, the researchers found no way to assess the combined effectiveness

of the interventions due to high heterogeneity in the type of the intervention and in the selected outcome measures.

Nørnberg and colleagues (2015) reviewed nudge interventions promoting vegetable consumption among adolescents. From the initial 2,158 findings in the publication databases, 12 studies were included in the final review. In their assessment, only one study was classified as strong quality as the other studies suffered from *"too small sample sizes and/or short intervention periods as well as the lack of information regarding the methodology in general, such as evidence of control of confounders and degrees of blinding"* (p. 4).

One of the most popular examples of nudge techniques is the positional influence on food choice. A prominent claim is that manipulating the proximity or the order of food can generate change in food selection and can promote healthier eating behavior. For example, putting salad to a more accessible position at a self-service food bar was reported to increase its sale (Rozin et al., 2011). In a recent systematic review, Bucher and colleagues (2016) included 18 studies in their empirical review from their initial 2,576 search findings. In the quality assessment, only one study received a positive rating because the other studies were found to suffer from poor levels of procedural description or questionable validity. This one study with higher quality, however, did not find significant effect of position on food choice.

The systematic review of Wilson and colleagues (2016) provided a similar picture of the empirical evidence for nudge techniques in the area of influencing food and beverage choices by salience and priming. Out of the 550 topic related articles, 13 were included in the review. As the majority of the articles received average or poor quality ratings and the results were substantially heterogeneous, the authors were unable to draw general conclusions about the empirical findings or to give recommendations for using these techniques for nudging.

In their scoping review focusing on diet-, physical activity-, alcohol-, and tobacco-related behaviors, Hollands and colleagues (2013) intended to map the nature of the existing evidence and to develop a conceptual map of the types of choice architecture interventions. After a systematic literature search, 346 articles were selected from their 804,919 search findings. They concluded that for *"the most part we were unable to indicate the direction or magnitude of the effects with any confidence"* as *"in most categories, studies were heterogeneous with respect to the populations, interventions, comparators or counterfactuals, outcomes and moderators assessed, and all studies remain to be examined in relation to potential confounders, the reporting of multiple outcome measures, subgroup effects and many other aspects of design, analysis, and reporting"* (p. 24).

In summary, these domain-specific systematic reviews aimed to provide answers to specific research questions targeted by nudge studies. However, they could not reach affirmative conclusions due to a number of deficiencies in the current empirical endeavors. The domain-specific scoping review of Hollands and colleagues (2013) further highlighted that these deficiencies can hinder the development of a solid evidence base concerning when and to what extent choice architecture interventions can be successfully applied.

In this paper, we present a domain-general scoping review of the nudge movement. First, we attempt to give an overview of the characteristics of the interventions and experimental designs employed in the nudge literature. Second, we intend to identify the most important obstacles that can hinder the evidence accumulation of the field and we aim to reveal whether the deficiencies found in specific domains characterize the field. Third, based on this overview and the insights of related disciplines, we will convey several recommendations for future studies. It is important to note, that our review focuses on the current state of the nudge movement and does not aim to provide an exhaustive review of all intervention studies using similar techniques. More precisely, the focus of the present work is restricted to studies which were published after the seminal book of Thaler and Sunstein (2008) and explicitly used the terms 'nudge' or 'choice architecture' to describe the applied interventions (see Methods section for the full list of inclusion and exclusion criteria). Therefore, our findings may not generalize outside of the choice architecture literature or to studies which were published before 2008.

Methods

Literature search

The review process included a systematic literature search in three academic databases (Scopus, PubMed, and Pubpsych) with a predetermined search strategy. The key terms applied on search were "nudge" and "choice architecture" combined with the Boolean operator 'OR'. Additionally, the terms "method*", "empiric*", "procedure", and "design", combined with the Boolean operator 'OR' were added to narrow down the focus of the search process on empirical papers[51]. As a result, the search strategy entered into databases was: ((nudge OR choice-

[51]To ascertain that our search included the most empirical studies possible, we ran a search query using only the terms "nudge" or "choice architecture" in the target databases. We looked up how the empirical sections of a randomly selected 100 findings were named. The search terms listed above covered all the articles that had empirical results in this sample.

architecture) AND (method* OR empiric* OR procedure OR design). We applied a restriction on dates, only including studies between January 2008 and May 2016. Additionally, the reference lists of the featured systematic reviews were also searched and used as a source of data.

Screening procedure

Following the literature search, the screening procedure was based on predetermined criteria (Table 1) and consisted of two consecutive stages. First, an initial screening of the abstracts and the titles of the candidate studies was carried out. Then, only the studies which passed this first filtering were examined based on the full-text by at least two of the authors to select the relevant papers to this review. In case of a disagreement, they consulted the first author and made the final decision together.

Table 1 summarizes the detailed list of inclusion and exclusion criteria applied in the screening procedure. We considered only English language empirical studies examining behavioral intervention techniques which were attributed as nudge by the authors or were connected to the choice architecture literature. We only considered behavioral interventions, in which participants' decisions were measured in a real choice situation or people's intentions or preferences were assessed in a hypothetical choice situation (e.g., I prefer A over B; I would choose A over B)[52]. Studies which exclusively focused on measuring people's attitudes or preferences towards objects or events were excluded. We applied no restriction on the population or the context of the study; however, studies restricting the freedom of choice and studies that used economic incentives, education, complex decision support systems or consultation as a nudge were excluded. The review focused only on peer-review articles, but further quality rating was not applied.

[52] We differentiated between hypothetical and real choice situations based on whether the participants' decisions had real consequences (such as spending money) or not (such as when in computer simulations participants are asked to imagine buying something).

Table 1

Inclusion and Exclusion Criteria

Inclusion Criteria
(1) Full-text papers.
(2) The paper's language was English.
(3) The paper was published in a peer-reviewed journal.
(4) The studies in the paper empirically investigated one or more behavioral intervention technique that was attributed as a nudge or was connected to the choice architecture literature by the original authors.
(5) The studies in the paper had behavioral outcome variables (not preferences or attitudes).
Exclusion
(1) Review articles, conference abstracts, and conference papers.
(2) The studies in the paper applied interventions that restrict the freedom of choice of the target population, included significant economic incentives or used education, complex decision support systems or consultation as a nudge.

Coding/data extraction

From the final list of target articles, one of the authors extracted the data and another author reviewed it. Table 2 summarizes the extracted data and the description of the labels. The categorization of intervention type was carried out independently by two authors. In case of any disagreement, another author was involved to resolve the issue by discussion. The summary of each study was sent by email to the corresponding author of the given study and the provided feedback was incorporated into our database. Subsequently, in the revision process, the categories ‘Linkage to theory’, ‘Emphasis on theory refinement’, ‘Food focus’ and ‘Power analysis’ were added.

Table 2

Label Structure of the Extracted Data and the Description of the Labels

Label	Description
Source	Reference to the study.
Study No.	Number of the study (if the article contains multiple studies), only those studies were included which were relevant to the context of the review.
Year	Year of publication.
Domain	Seven categories were predefined based on the review of the relevant literature (transportation, sustainability, prosocial behavior, health, finance, education, consumer choice, other).
Subdomain	A short summary of the aim of the intervention indicating the specific domain (e.g., making healthier food choices; studying more).
Category of intervention	Based on the taxonomy of Münscher, Vetter, & Scheuerle (2015) the interventions were sorted into nine categories, and in case a study implemented multiple different choice-architecture techniques all of them were listed. The categories were the following: translate information (e.g. framing techniques, simplification of information), make information visible (e. g. using feedback techniques, presenting external information), provide social reference point (e. g. referring to a descriptive norm or to an opinion leader), change choice defaults (e.g. opt-in, opt-out

	techniques or prompted choice), change option-related effort (e.g. financial or physical effort), change range or composition of options (e.g. changing categories or grouping of options), change option consequences (e.g. changing social consequences of the decision or connecting it to benefit or cost), provide reminders (e.g. making information more salient or easier to access), facilitate commitment (e.g. supporting self or public commitment to choice).
Location	The country in which the intervention was carried out or, in case experiments where the location is not relevant (e.g., online experiments), the country from which the sample was collected, if known.
N (after exclusion)	Number of units in the sample that were used for the analysis in the study.
Number of nudge intervention tests	Number of all intervention groups in which there was a nudge intervention, for example, in a between-subject study, if there was a control, intervention A, intervention B, and a mixed intervention AB group, the number of nudge intervention tests was determined three. The term 'non-conclusive' was applied if two authors could not reach a decision on the number due to the design of the study.
Number of failed nudge interventions	The number of those intervention groups in which the nudge intervention showed no significant effect, including the number of those intervention groups in which the effect was in the opposite direction to the desired direction specified by the authors; 'Non-conclusive' was used in case the study reported several output variables with mixed results and two independent co-authors could not reach a decision.

Moderators	List of variables moderating the effectiveness of the nudge that was found significant.
Linkage to theory	If models or theories are mentioned in the paper which specify the hypothesized causal mechanisms or processes behind the intervention(s), the study was marked as (1), otherwise (0).
Main motivation of the study	(1) intervention assessment; (2) exploration of moderators or underlying processes; (3) other.
Emphasis on theoretical contribution	If either there is any suggestion in the abstract on what the results imply about the underlying causal processes or the *Main motivation of the study* was the "*exploration of moderators or the underlying processes*", the study was marked as (1), otherwise (0).
Pilot	Was there a pilot for the nudge intervention or interventions (1) or not (0)? If the study itself was a pilot, (0) was indicated.
Field/Lab	Field experiment (F), laboratory experiment with hypothetical choice situation (LH) or laboratory experiment with real choice situation (LR).
Follow-up	In case there was a follow-up, the time between the intervention and the follow-up periods.
Food focus	Does the intervention aim to change eating or drinking related behavior? (1) yes (0) no.
Power Analysis	Do the authors report performing power analysis before conducting the study to determine their target sample size? (1) yes (0) no.

Results

2670 entries were identified through the search in the three academic databases. 1714 articles were excluded based on the initial screening of the abstracts and the titles. 51 duplicates were also removed. The remaining 905 articles were downloaded from which 799 entries were excluded based on the predetermined criteria. The final list of target studies contained the remaining 106 entries and 10 articles were added from the screening of the reviews (See Figure 2.) That is, our final database contained 116 empirical articles with 156 studies and more than 422 tested interventions. Note that one article may contain more than one study and in one study more than one intervention can be tested. We conducted the analyses on the characteristics of the studies or the interventions, depending on the questions of interest.

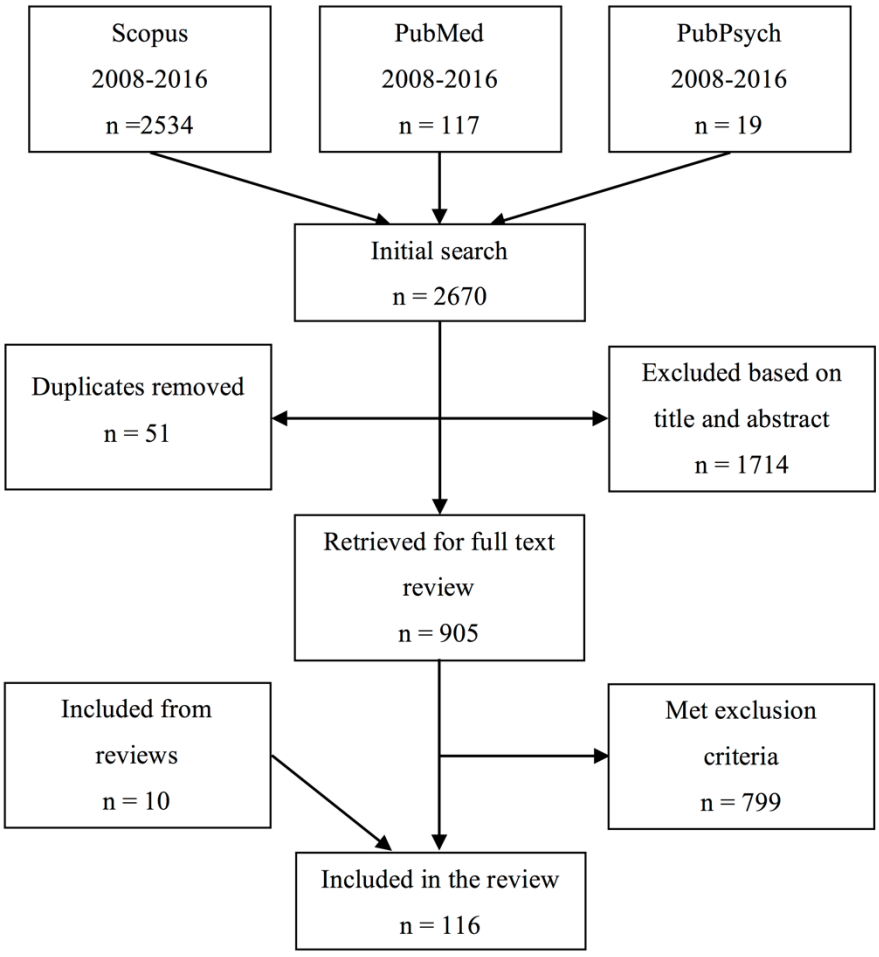

Figure 2. Information flow of the screening procedure

49% of the studies were conducted in the US and 38% in Europe predominantly in the United Kingdom, the Netherlands and Germany. Only 13% of the studies were run in countries from other continents (Australia, Bangladesh, Colombia, Israel, Japan, Kenya, Vietnam, Zimbabwe, Malawi). We observed an increment in the number of empirical choice architecture studies per year between 2008 and 2015, with a break in this growth in 2012 (Figure 1). While between 2008 and 2012, 62% of the studies were conducted in the US, between 2013-2016 the same number was 43%[53].

Fifty-four percent of the studies were field experiments, while 46% were conducted in the lab. In 75% of the lab studies, participants faced a hypothetical choice situation while in the rest the main outcome variable was a real choice. Descriptive statistics of the human sample sizes after exclusions are presented in Table 3. It is important to note that the data was heavily skewed. The median sample size was the smallest in lab studies with real choice, larger in lab studies with hypothetical choice, and the largest in field studies.

Table 3

Descriptive Statistics of the Sample Sizes by the Type of the Study

	M	*SD*	*Mdn*	*Min*	*Max*
Real choice lab studies	199.65	228.36	150	65	1000
Hypothetical choice lab studies	446.35	702.34	198.5	29	2802
Field studies	3003.06	7732.67	734	18	54385

Note. Only studies with human samples were considered.

We observed that the interventions in each study were linked to certain theories or models. (See Table 2. for the definitions of categories). Twenty-four percent of the studies focused on the exploration of moderators or underlying processes, 74% of the studies were mainly motivated by assessing the applicability of the examined interventions in a given setting, while 2% had some other goal. We applied an additional categorization to determine whether an emphasis of the papers was on theoretical contribution. If there was any suggestion in the abstract on what the results imply about the underlying causal processes or theories, or if the

[53] This change might reflect both the spread of the nudge movement and the fact that it was initiated by two prominent researchers from the US.

main motivation of the study was the exploration of moderators or the underlying processes, the study was marked as having an emphasis on theoretical contribution. We found that in 42% of the studies fell into this category.

If either there is any suggestion in the abstract on what the results imply about the underlying causal processes or the *Main motivation of the study* was the "*exploration of moderators or the underlying processes*", the study was marked as (1), otherwise (0).

The interventions of interest were piloted in 13% of the studies. The authors used power analysis to determine the sample size in 7% of the studies. Seven percent of the studies included follow-ups to measure whether the effect of the intervention sustained over time.

Ninety-three percent of the studies in our database contained at least one successful intervention, yet only 18% of the reported interventions were unsuccessful. Forty-seven unique variables were found to moderate the effectiveness of the nudges of which findings 26 were based on correlational and 21 on experimental design.

The domain of health was the most studied field in our sample with 67 nudge studies (42%) undertaken. Notably, 64% of the health-related studies aimed to change eating or drinking related behavior. The second most studied domain was sustainability (19%), which was followed by consumer choice (10%), prosocial behavior (8%), finance (9%), transportation (6%), and education (4%). Table 4 summarizes the distribution of the studies identified in the review by intervention and category.

Table 4

Number of Studies by Category of Intervention and Domain

	Domain								
Category of intervention	Consumer choice	Education	Finance	Health	Other	Prosocial behavior	Sustainability	Transportation	**Total**
Change choice defaults	5	2	8	18	1	8	13	2	**57**
Change option consequences	-	-	1	4	-	2	-	-	**7**
Change option-related effort	-	-	1	16	-	-	1	-	**18**
Change range or composition of options	4	-	-	6	5	-	1	-	**16**
Facilitate commitment	-	-	-	2	1	3	3	-	**9**
Make information visible	-	1	6	19	-	1	7	2	**36**
Provide reminders	3	3	1	20	1	3	2	1	**34**
Provide social reference point	2	-	1	9	-	1	6	2	**21**
Translate information	6	1	-	11	-	-	3	4	**25**

Note. Studies containing interventions from different intervention categories are represented in all the corresponding rows. As a consequence, the total number of studies in Table 4 exceeds the number of independent studies in our sample.

Table 5 provides an overview of each intervention category. Consider, for example, *changing choice defaults*, the most studied intervention type, which was applied in 57 studies with a great variety of settings, such as choosing energy efficient lightbulb or increasing loan statement. Throughout the studies, 18 variables were found to significantly moderate the effect of the default intervention; out of which 8 were experimental manipulations (e.g., financial cost, value of the default, social reference point), while 10 were explored using correlational design (e.g., self-control, state, expertise, future repayment concern). Other intervention types were investigated less frequently. For instance, only 7 studies used nudges involving *changing option consequences* and only 1 significant moderator was explored (making a pledge). Note, however, that our table does not provide a complete picture of the empirical basis behind each intervention type, but delivers a preliminary summary in the light of the recent advancement on the field of nudging.

For the annotated bibliography of this dataset see Szaszi, Palinkas, Palfi, Szollosi, and Aczel (2016) which contains a summary of the target population, the desired behavior, the interventions used and the result of each study from this review.

Table 5

Overview of the Intervention Categories Describing the Subdomains and the Moderators Associated with each Intervention Category

Intervention	**Definition**	**Context / Subdomain**	**Moderators**
Change option consequences	Linking micro-incentives to the consequences of the options.	Carrying epinephrine for allergy, Increasing physical activity, Shortening the time of the peer review of an article, Encouraging participation in chlamydia screenings, Making healthier food choices, Donating to charity,	Making a pledge (correlational)
Facilitate commitment	Having participants committed to an option.	Re-using towels, Recovering from nail-biting, Saving energy (2), Discouraging dishonesty (2), Enhancing the appropriate prescription of antibiotics, Donating to charity, Increasing vaccination	Specificity of commitment, Value of the default (experimental), Preferring immediate gains, Goal setting, Making a pledge (correlational)
Change range or composition of options	Making changes in the alternative options presented.	Donating to carbon offsetting programs, Making healthier food choices (4), Modifying product preferences, Increasing plant-based food choices, Modifying food choices, Modifying food and beverage choices (2), Reducing gender bias, Choosing more cost effective health insurance (6)	Financial cost (experimental)
Change option-related effort	Marginally changing the amount of effort the selection of an option requires.	Making healthier food choices (6), Making healthier food and beverage choices (2), Drinking more, Decreasing food intake (2), Increasing low-calorie meal choices (2), Increasing low-calorie snack choices, Increasing physical activity, Making better health insurance decisions, Increasing recycling performance, Increasing redemption	-

Provide social reference point	Providing information on other people's behavior.	Making healthier food choices (3), Contributing to public goods, Choosing environmentally friendly vehicles, Decreasing red light crossing behavior, Choosing environmentally friendly products, Drinking more, Modifying food choices (2), Seeking alcohol-related health information, Increasing the choice of chemotherapy (2), Increasing vaccination, Increasing sustainable consumption (2), Reducing pollution(2), Increasing redemption, Reducing in water consumption, Decreasing energy consumption	Framing the social reference point, Incremental Benefit (2), Social reference point, Value of the default, Identified or unidentified neighbors, same or different neighborhood (experimental)
Translate information	Changing the format or presentation of information without changing the content.	Choosing automatic prescription refills (2), Choosing environmentally friendly products, Choosing environmentally friendly vehicles (3), Decreasing red light crossing behavior, Donating to an environment saving organization, Encouraging participation in chlamydia screenings, Increasing plant-based food choices, Increasing vaccination (2), Making appointment for dental check-ups, Making healthier food choices (5), Modifying genetically modified vegetable choices, Reducing calorie intake, Reducing prescribing errors, Selecting healthy food products, Saving energy, Studying more, Taking the stairs instead of using the elevator	Gender, Country, Dieting, Executive function strength, Psychological closeness to future generations, Children in the family (correlational), Motivation, Type of the product - healthy or unhealthy, Framing the social reference point, Price of the product (experimental)

Provide reminders	Modifying the salience and the ease of access of options.	Increasing handwashing, Selecting hotels, Drinking more (2), Making healthier food choices (5), Recovering from nail-biting, Increasing physical activity, Increasing vaccination (2), Reducing procrastination, Making appointment for dental check-ups, Shortening the time of the peer review of an article, Increasing fruit and vegetable choices, Increasing plant-based food choices, Increasing the frequency of breast cancer self-check, Reducing prescribing errors, Enhancing the appropriate prescription of antibiotics, Increasing the amount of restitution paid, Saving energy, Increasing whole wheat bread roll choices, Increasing lower-calorie milk choices, Modifying food choices, Increasing redemption, Increasing college enrollment, Donating to charity, Decreasing food intake, Switching the lights off, Remaining enrolled at university, Decreasing road accidents, Selecting healthy food products	Preferring immediate gains, Response to the text messages, Timing of prompts, Class preferences, GPA, Free student help enrollment, Making a pledge (correlational), Message content, image type (experimental)
Make information visible	Making visible information that is not apparent or readily available.	Carrying epinephrine for allergy, Choosing environmentally friendly routes, Decreasing red light crossing behavior, Making healthier food choices (7), Drinking more, Increasing physical activity, Increasing loan repayment (3), Increasing environmentally friendly behavior, Making better health insurance decisions, Making healthier food and beverage	Future repayment concern (2), Propensity to pay, Belief in global warming, Executive function strength, Children in the family (correlational),

		choices (3), Modifying real estate choices (2), Reducing calorie intake, Increasing recycling performance, Increasing plant-based food choices, Increasing the amount of restitution paid, Saving energy, Increasing low-calorie meal choices (2), Taking the stairs instead of using the elevator, Reducing water consumption, Deploying low tidal volume ventilation, Remaining enrolled at university, Choosing more cost effective health insurance (4), Selecting healthy food products	Framing the social reference point (experimental),
Change choice defaults	Changing which options are pre-selected and what needs to be chosen actively.	Donating to carbon offsetting programs, Choosing therapy for Autism Spectrum Disorder, Contributing to public goods, Choosing energy efficient lightbulbs (3), Making investments (2), Modifying parking preferences (2), Achieving safer sedation doses, Increasing vaccination (3), Choosing automatic prescription refills, Increasing college enrollment, Saving energy (4), Reducing innovation rejection (2), Increasing loan repayment (4), Choosing relevant medical tests, Increasing blood donation, Increasing protein consumption, Making healthier food choices (5), Modifying therapy decisions about cancer treatment, Shortening the time of the peer review of an article, Modifying prescription choices, Increasing attendance to a business training program, Donating to charity (3), Donating to Greenpeace, Modifying tip	Financial cost, Value of the default (2), Social reference point, Consistency between the focus of the default option and recently collected arguments, Ego-depletion, Format of presentation, Donor status, Cooperation strategy (experimental) Self-control, State, Expertise, Future repayment concern (2), Propensity to pay, Awareness, Type of social media activity, Education, Bar or restaurant,

	choices, Reducing prescribing errors, Increasing organ donation, Reducing medical costs, Modifying wine-sales, Decreasing alcohol consumption, Decreasing food intake, Deploying low tidal volume ventilation, Increasing HIV testing, Choosing more cost effective health insurance, Increasing sustainability in engineering, Reusing towels (2), Switching to electronic bank statements, Switching to electronic bill statements	Environmental consciousness (correlational)

Note. The underlying database (Supplementary materials) can be used to search for further information on the studies mentioned in this table. Any label (domain, intervention, subdomain etc.) described in Table 2 can be used as a filter to identify the studies of interest.

Discussion

The first aim of our scoping review was to investigate the current state of the nudge movement. Our findings confirmed that the empirical basis of the movement is dynamically growing. The applied nature of the field is reflected both in the high proportion of field studies and in the rising number of application of nudge techniques in real-world settings. The most intensive academic research activity is concentrated in the US, but research groups have started to investigate nudges on all the five continents, with Europe being the second most active of all and with a growing contribution.

The spreading of the nudge movement is also apparent in the variety of the investigated domains and intervention types. Although health related research dominates the movement with a special focus on eating and drinking related behavior, a considerable number of empirical investigations have been carried out in other areas such as sustainability or consumer choice. The diversity of subdomains present in choice architecture studies is inspiring: interventions have been tested in 87 subdomains, reflecting the wide applicability of the nudge techniques. The studies showed high heterogeneity along numerous dimensions such as target population, intervention type, target behavior, experimental design, the way of reporting and the type and number of outcome measures. A great number of factors moderating the effectiveness of the nudge techniques has also been revealed across the different domains and experimental designs.

Policy makers have often relied on nudge-like techniques in the past to influence human behavior, but due to lack of rigorous research, these attempts were mostly based on the pure luck of trial-and-error. The nudge movement aimed to take the next step and provide an evidence-base to practitioners in their attempts to promote socially advantageous behavior. However, previous domain-specific nudge reviews suggested that, for identifiable reasons, the field is greatly limited in its ability to accumulate evidence, and to predict when and to explain why nudges work.

In the present paper, we also explored some characteristics of the choice architecture literature which could potentially affect the process of evidence accumulation and found that the nudge movement has a considerable room for improvement in several respects. In what follows, we explicate these problematic characteristics and propose some possible steps which could support the improvement of the field.

Two more things need to be clarified here. First, neither the problems nor the solutions are necessarily unique to the nudge movement. By highlighting them we join to the many researchers who urged to make changes in order to support the evidence accumulation in behavioral science (e.g., Michie, Johnston, Francis, Hardeman, & Eccles, 2008; Munafó et al., 2017; Simmons, Nelson, & Simonsohn, 2011). Nevertheless, to our best knowledge, the present paper is the first work exploring the need for such changes in the field of nudging and our analysis suggests that many problems are present despite being praised elsewhere.

Second, as we are not the first identifying such problems and offering solutions, we proudly use the knowledge accumulated before us. That is, the identification of the hurdles of evidence accumulation are rooted in our findings, but they are also supplemented with the concerns raised in previous domain-specific nudge reviews and with related considerations raised in other fields (such as the broader field of behavioral interventions or psychological science). When summarizing the solutions, we aimed to synthesize the relevant recommendations from other fields and apply it to the characteristics of the problems and the nudge movement.

Hindrance to evidence accumulation

Nomenclatures

The fact that different categorizations (nomenclatures, taxonomies, labels) of choice architecture techniques are used in parallel in the literature poses several challenges for choice architecture researchers: it makes the accurate replication and implementation of interventions as well as the creation of systematic reviews harder (Abraham & Michie, 2008; Münscher et al., 2015; Michie et al, 2013). The root of these challenges is multifaceted. First, different nomenclatures use different basis for their categorization: they focus either on the underlying cognitive processes (e.g., Dolan et al., 2012; Datta & Mullainathan, 2014) or on the used interventions (e.g., Johnson et al., 2012; Münscher et al., 2015). Second, the categories used in the current nomenclatures are often redundant and non-exhaustive: that is, one label can refer to several intervention techniques and the list of labels does not cover all forms of nudge interventions.

Experimental and reporting practices

Previous systematic reviews of the choice architecture studies found insufficiencies both in the quality of the design and the reporting of the studies. For example, these reviews found that the procedural level description of the interventions was often not detailed enough for proper

replication or that given the insufficient quality of reporting of complex interventions and their interacting components, it was not possible to disentangle the effect of these components (for a more detailed discussion see Skov et al., 2013; Nørnberg et al., 2015; Bucher et al. 2016). Although in our scoping review we did not systematically review the quality of the studies, we have found similar patterns which made it challenging to condense the results.

We also observed that only 6% of the studies estimated the necessary samples size before data-collection. For a comparison, 95% of the articles in medicine (Charles, Giraudeau, Decharrtres, Baron, & Ravaud; 2009) reported sample size calculations. Although, there have been several reporting guidelines developed in the last decades providing checklists to support and improve the quality of reporting. (e.g., Chan et al., 2013; Begg et al., 1996; Moher, Schulz, & Altman, 2001; Schulz, Altman, & Moher, 2010; Hoffmann et al., 2014), we found that in our sample only in 3 papers was any kind of reporting guideline used.

Publication and reporting bias can distort the estimated effect sizes of the interventions, slow down the exploration of boundary conditions of the nudges, and hinder the falsification of hypotheses explaining the underlying mechanisms (Rosenthal, 1979; Ferguson & Heene, 2012)[54]. As the prevalence of publication and reporting bias can be mitigated by the application of preregistration systems, the preregistration of experiments is encouraged in a growing number of journals (Lindsay, 2016; Nosek & Lakens, 2014) and this practice has become the standard for randomized controlled trials in clinical medicine (Lenzer, Hoffman, Furberg, & Ioannidis, 2013). In contrast, we found that not a single study in our sample was mentioned to be preregistered.

Connections with theories

It has been shown that interventions connected to theories[55] are more effective relative to those which are not (e.g., Webb, Joseph, Yardley, & Michie, 2012; Painter, Borba, Hynes, Mays, & Glanz, 2008). We also found that in each study, the applied interventions were connected to some choice architecture theories or frameworks (e.g., Smith, Goldstein, & Johnson, 2013; Vlaev

[54]We observed that the 83% of the reported choice architecture interventions were successful (i.e., statistically significant) while based on previous estimates one would expect the proportion of significant studies to be around 20% considering that the average power in psychology has been estimated to be 35% (Bakker, van Dijk, & Wicherts, 2012), using the standard 5% alpha criterion and assuming an equal proportion of true and false hypotheses.

[55]By theory we mean an "integrated summary of the hypothesized causal processes" (p. 662, Michie et al., 2008) which make predictions about when, to what extent, and why the intervention works.

& Dolan, 2015; Dolan et al, 2012). However, our findings also point to an important characteristic of the literature: the main motivation of the studies was overwhelmingly the assessment of the intervention effectiveness in a given setting, and in only fewer than the half of the studies was the emphasis on the development of theories and the deeper understanding of the causal processes.

Recommendations

Developing and Reaching Consensus on the Nomenclature

We urge the further development of the choice architecture taxonomies since without a comprehensive nomenclature with clearly labeled, non-overlapping and non-redundant categories (for an example see Michie et al., 2013) the necessary consensus can hardly be reached. Here, we make several suggestions on the necessary steps of the taxonomy development. First, the relationship between cognitive processes and intervention techniques is complex. As on most interventions we are yet not well formed about the underlying processes (Münscher et al., 2015), we think that a consensus could be more easily reached by using intervention techniques as a structuring principle (See also Münscher at al., 2015, Michie et al., 2013).

Second, as the current intervention-based taxonomies do not provide an exhaustive list of categories and subcategories to cover all the various forms of nudge intervention taxonomies, they need to be further developed. This kind of continuous development would not be unprecedented and could lead to a point of consensus. Consider, for example, the evolution of the taxonomies in the field of behavioral interventions. The first systematically developed taxonomy contained only 22 behavior change techniques (Abraham & Michie, 2008). However, it was further advanced in a Delphi-type expert consensus process. As a result, the most recent nomenclature describes 93 hierarchically clustered techniques (Michie et al., 2013) and was widely applied since its publication[56].

Third, as nudges and other behavioral interventions (such as trainings or financial incentives) often target similar problems, we suggest that the nudge and behavioral intervention taxonomies should be integrated. This step could support the accumulation of the scientific evidence from more and broader domains and could also help the comparability of the

[56] The article was cited more than 750 times as of January 2017.

interventions regardless whether they are, by definition, nudge or non-nudge (Hansen, Skov, & Skov, 2016; Hollands et al., 2013).

Finally, to foster the understanding of the effect of the delivery mode (e.g., in-person, digital, printed, etc.) of choice architecture interventions, we suggest to adopt and apply the taxonomy of the mode of deliveries (Carey, Evans, Horan, Johnston, West, & Michie, 2016). Although the work of Carey and colleagues was developed in the broader field of behavioral change interventions, its categories can be applied also to choice architecture based interventions.

Improving experimental and reporting practices

We also recommend the usage of reporting guidelines. It has been shown in other fields that the adoption of such guidelines can improve the quality of reporting in the scientific literature (Moher, Schulz, Simera, & Altman, 2010). The most widely used guideline is the Consolidated Standards of Reporting Trials (CONSORT) statement which includes a 25 item checklist on how to report parallel group randomized trials (Schulz, Altman & Moher, 2010). The Template for Intervention Description and Replication (TIDieR) is an extension of CONSORT and provides a checklist on how properly describe interventions (Hoffman et al, 2014). The CONSORT-SPI is another CONSORT extension under development, which is specifically suited for social and psychological interventions (Grant, Mayo-Wilson, Melendez-Torres & Montgomery, 2013). Note, beyond these three, there are many other specific reporting guidelines developed and that the choice of which guideline to adopt should be dependent on the specific research design applied[57]. However, the above mentioned three guidelines can provide a starting point for researchers running and reporting choice architecture interventions.

Along many others (e.g., Munafo et al., 2017; Gonzales & Cunningham, 2015), we also recommend the usage of the public preregistration systems as these can mitigate publication and reporting bias. There are several free and easily applicable solutions (e.g., Open Science Framework, American Economic Association's registry, Aspredicted.org, etc.).

[57]For a collection on such guidelines see http://www.equator-network.org/.

Learn more about why, when and to what extent interventions work

We argue that researchers and practitioners, not focusing on theoretical advancement but on the assessment of the applicability of choice architecture interventions, could also contribute more to the evidence accumulation. This can be achieved, for example, by using the taxonomies and reporting standards of the broader academic community, and by putting more focus on the key variables, which may moderate the effectiveness of the nudge in their setting. We suggest that choice architects should identify, and if possible, assess the effect of these factors. The reporting of these critical variables could contribute to the understanding of the causal pathway behind the interventions. Alternatively, if more academic choice architecture studies focused on testing the generalizability and the predictions of theories formulated in previous research it could further support the refinement and advancement of those theories (see also, Davis et al., 2015; Michie & Johnston, 2012; Oakley et al., 2006).

Furthermore, exploiting the growing number of empirical evidence, domain-general intervention-specific systematic reviews and meta-analyses should be conducted as these could support the advancement of theories behind the interventions. Currently the number of this type of work is still scarce. Although the many available domain-specific narrative reviews are convenient tools to showcase the applicability of the choice architecture interventions, but less useful contributors to the understanding of why, when and to what extent interventions work.

Limitations

The present review is not based on an exhaustive list of studies which apply nudge-like interventions as our search strategy was restricted to studies which are connected to the choice architecture literature and were published after 2008. Therefore, our findings may not generalize outside of the choice architecture literature. Additionally, due to the limitation of the used academic databases, our review may also miss empirical findings from the nudge literature. However, we think that the present work yields a good estimate of the characteristics of the nudge movement and can be used to map the available evidence.

Conclusion

The choice architecture movement has presented a wide variety of interventions throughout the last several years. The low resource demand and broad applicability of these interventions

invoked widespread interest and convinced policy makers to use these tools to tackle important societal problems. However, our review suggests that, for identifiable reasons, the field is greatly limited in its ability to provide process level explanation of these interventions and to summarize their boundary conditions, therefore, the effectiveness of the different interventions across different domains of applications cannot be predicted. In other words, by now, the choice architecture movement provided tools for the practitioners without accompanying instructions for how and when to use the different intervention techniques. A desirable next step of the movement would be to eliminate the obstacles of evidence accumulation and by that to create the necessary circumstances for the development of theoretical frameworks. These frameworks could provide predictions for the effective use of the different nudges and could promote the development of new choice architecture techniques. In the present review, we aimed to highlight these obstacles and to make recommendations on how to overcome them.

Data and Analysis Code

The analysis code and the underlying databases are openly available at https://osf.io/dxbtz/.

Acknowledgment

We would like to thank Marton Kovacs, Mark Zrubka and Boglarka Nagy for their assistance, Melissa Wood for proofreading the manuscript and Melinda Szászi-Szrenka for her supporting love and patience throughout the study. Aba Szollosi was supported by the "Nemzet Fiatal Tehetségeiért" Scholarship (NTP-NFTÖ-16-1184).

General Discussion

The present dissertation focused on the factors influencing and the cognitive processes underlying human decision-making. In the first two studies, we explored how individuals solve the Cognitive Reflection Test and the denominator neglect task. By tracking the thought processes, we tested the time-course assumptions of the default-interventionist theory, a central decision-making theory discussing the influencing role of the first thoughts and the lack of intervening analytic thinking behind many reasoning failure. The results of Study 1 and 2 suggested that the default-interventionist model needs some refinement. In Study 3, we created a novel task (the Guided Movement Task) and investigated the process by which action dynamics influence individuals' decisions. The findings provided evidence that the movements influence the cost of the change but not the value of the options. In the last study, we conducted a systematic review of choice architecture interventions. These interventions use the insights from the psychology literature on the factors influencing decision-making. Based on the results of the systematic scoping review, we formulated several recommendations on how scientific practices should be improved in order to enhance the process of evidence accumulation on when and why nudges work and on which nudge to use in a given situation.

In the next section, I summarize the main findings from each study, which is followed by the discussion of the limitations. I close the dissertation by delineating the possible future directions of my work[58].

Summary and main findings of the studies:

In Study 1 and Study 2, we investigated the assumptions of the default-interventionist theory in the Cognitive Reflection Task (CRT) and the denominator neglect task by examining how the initial incorrect responses influence individuals' decisions. The default-interventionist theory assumes that many of the decision biases happen because individuals produce an incorrect

[58] It is said that the current replication crisis is partly driven by researchers overclaiming their results. Accordingly, although I think it is always essential to keep in mind how a given study fits into the big scientific scene of psychology, I find it important to explicitly articulate that I deliberately aimed to keep the discussions focused and not overgeneralize the results of my studies.

heuristic response and fail to override this first response with further deliberation. Many theories explaining how individual differences arise in situations where biased decision making is often observed also incorporated the default-interventionist assumptions (e.g., Stanovich & West, 2008). Using process-tracing techniques, we asked whether the default-interventionist assumptions hold or this view needs some further refinement. Do individuals sometimes have a correct first response in these situations and if yes, how often? Do individuals sometimes reflect about their first incorrect response but still fail provide the correct answers? Do individuals sometimes change their mind after an initial correct thought? And why individual differences arise in these tasks: Are higher capacity people more likely to start their thinking with a correct initial answer? Are they more likely to stay with their initial answer when it is correct? Are they more likely to change their mind when their initial answer is incorrect?

In Study 1, protocol analysis was conducted: participants came to the laboratory and were asked to solve the three items of the CRT whilst thinking aloud. Furthermore, participants completed several individual differences measures assessing reflectivity and cognitive abilities including the AOT (Baron, 1993), REI (Pacini & Epstein, 1999), BBS (De Neys, Moyens, & Vansteenwegen, 2010), BNT (Cokely et al. 2012), SIs (Mata, Schubert, & Ferreira, 2014) and the BIDR (Paulhus, 1991). The protocol analysis revealed that in 77% of the cases when reasoners gave the correct answer, reasoners started their thinking with a correct response or with a line of thought leading to the correct answer. We observed that even when individuals provided an incorrect answer on the tasks of the CRT, they reflected about their response in 39% of the cases. We have also found that both the numeracy measures and the reflectivity measures predicted the performance on the CRT. However, we did not find evidence, that individuals cognitive capacity predicted the correctness of the first responses.

Study 2 consisted of a mouse-tracking experiment and its replication. Participants came to the computer test room and after a test phase, responded to 54 trials of the denominator neglect task. The respondents' mouse movements were recorded during the experiment. After this laboratory-based offline session, participants filled out two cognitive capacity measures: the BNT and an IQ test (Kovacs et al., 2016). Our mouse tracking analysis revealed, that individuals initially moved the mouse cursor towards the correct answer option in 43% and 45% of the incongruent trials across the two experiments. We also observed that in around 3% of the cases, individuals started with a commitment towards the correct response but changed their mind and finally chose

the incorrect answer. The analysis of individual differences provided further interesting insights: similarly to Study 1, we did not find evidence that high capacity individuals produce correct initial answers more frequently. However, our analysis revealed that compared to low capacity reasoners, high capacity individuals revise their first answer more frequently if it is incorrect and make fewer changes if it is correct.

The results of Study 1 and Study 2 indicate that the default-interventionist view and the theories building on the default interventionist view need to be refined. The findings suggest that decision biases often don't arise because individuals fail to engage in analytical reasoning after an incorrect initial response. We observed that individuals sometimes have a correct initial response and more importantly, in a small amount of cases they change their mind and provide an incorrect final answer. That is, sometimes decision biases arise even if someone has a correct first thoughts. Second, in some cases, individuals fail to make the correct decisions even if they engage in analytical reasoning[59] after the initial incorrect response. This can happen for several reasons: sometimes people lack the necessary declarative knowledge and strategic rules (mindware, Perkins, 1995) that are needed to solve the problems, while other times when individuals try to make a decision, instead of collecting all the information necessary to make a well-informed decision, people use reflective reasoning to rationalize or justify their first thoughts (Evans, 1996; Evans & Ball, 2010; Wason & Evans, 1975). That is, sometimes decision biases arise even if someone engage in analytical reasoning after an incorrect initial response. Our findings supported the default-interventionist-based view of individual differences: we found that individual differences in cognitive capacity did not predict the correctness of the initial response instead individual differences arose after the first response was formulated. We observed that this can happen for two reasons: high capacity reasoners more frequently changed their mind when their initial response was incorrect, but less often changed their mind when their initial response was correct.

In Study 3, we designed a novel task to disentangle the influence of the continuous act of getting closer to a choice option from the increased motion cost associated with changing the movement direction to another option. In two laboratory-based experiments, participants were

[59]Probably this is not a finding that anyone should be surprised. What is more surprising is that only a limited number of studies investigated the extent to which the default-interventionist pattern holds.

presented with 88 and 40 experimental trials showing pairs of human faces and risky choice tasks, respectively. The respondents were asked to choose the face which they found more attractive or the risky options which they preferred. Participants had to indicate their answer with the mouse cursor in a Guided Movement Task which directed the respondents' computer mouse movements by constraining the movement space. The results suggested that only the cost associated with the motion influenced the individuals' decisions and moving towards an option itself did not have an effect on the likelihood of choosing one option over the other. The Bayes factor analysis indicated support for the absence of this latter effect. These results were identical in both of the experiments involving faces or risky choice tasks. The findings provide further evidence that action execution is a key part of the decision-making process and can have a significant influence on the decisions. The GMT allows us to conclude that the effect of action dynamics on the decision is probably driven by the cost associated with the movements and not by other cognitive processes such as increased commitment effect towards one of the options.

In sum, the results of Study 1, Study 2 and Study 3 deepen our understanding on the factors influencing and cognitive processes underlying human decision-making, provide information on why and when decision biases arise and suggest some refinement to the default-interventionist-based theories and the embodied choice theory.

In Study 4, I extended my scope and focused on the nudge interventions which build on the insights gained from psychological studies examining the key factors influencing decisions. My colleagues and I reviewed 422 interventions from 156 studies to create a domain-general scoping review of the choice architecture movement. We conducted a predefined search strategy in three academic databases to identify all the nudge studies after 2008. Subsequently, we checked each study whether it meets our inclusion criteria. The results showed that the empirical basis of the movement is dynamically growing. Interventions have been tested in 87 subdomains, reflecting the wide applicability of the nudge techniques. However, we found that the field is limited in its ability to accumulate evidence which is necessary to predict when and why nudges work: there are insufficiencies in both the quality of the design and the reporting of the studies. Twenty-four percent of the studies focused on the exploration of moderators or underlying processes, 74% of the studies were mainly motivated by assessing the applicability of the examined interventions in a given setting, while 2% had some other goal. We also found that 2% of the studies used guidelines aiming to improve the quality of reporting, no study in our database was preregistered,

and the intervention nomenclatures were non-exhaustive and often have overlapping categories. Accordingly, we urge the further development of choice architecture taxonomies using the intervention techniques as a structuring principle. We also recommend using reporting guidelines to increase the quality of the reports as well as the preregistration of the studies. Last but not least, we argue that the process level explanation of these interventions and their boundary conditions is greatly limited and researchers should focus more on exploration of underlying mechanisms and moderators to develop models predicting when, why and to what extent different nudge interventions work.

Limitations

Although the limitations of each empirical study have been discussed in detail, I would like to provide a summary of the main limitations of my studies. First, a core question of Study 1 and Study 2, is whether we could correctly identify the first thought which come to the individuals' mind. Although the reaction times in both studies suggest that we could, we cannot unequivocally exclude that some of the participants did not articulate their first thought in the protocol analysis or did not move the mouse cursor in the mouse-tracking study toward the option which they were first committed.

Second, former studies have shown that there is a great inter- and intra-individual variability across different kind of problems and decision situations (Aczel, Bago, Szollosi, Foldes, & Lukacs, 2015). Future studies need to explore whether our findings generalize to other tasks and situations where decision biases arise or where embodied choice may play a role (Study 1, 2 and 3).

Third, I have used several self-administered, survey-type, questionnaires (e.g., AOT, REI in Study 1). Such self-reported questionnaires are sensible to social desirability bias and also to short-term recall biases. Similarly, the method of protocol analysis is also exposed to the danger caused by social desirability.

Fourth, it has to be noted, that some studies argue that cognitive capacity is a multi-dimensional construct. However, given the questionnaires we used, we cannot make conclusions on which component of cognitive capacity caused the observed effects (Study 2).

Fifth, although Study 2 had some experimental elements, the statistical results of Study 1 and Study 2 are based on a correlational design with all the limitations of such approach.

Finally, I wanted to highlight what the studies (especially Study 1 and 2) were aiming not to achieve. We did not test whether the responses were generated by qualitatively different systems (cf. Type 1 and Type 2), or by one single system (Keren & Schul, 2009). Although dual-process theories have become increasingly popular in the recent years (e.g., Chaiken & Trope, 1999; Sloman, 1996) and many of the default-interventionist theories also build on a dual-process framework (Stanovich & Evans, 2013), we only tested the assumptions of the default-interventionist view regarding the temporal-course of thinking. We did not test either the features of the responses: whether the first answer was automatic, controlled or unconscious. Consequently, we cannot add insights to the single- vs. dual-process discussion.

Future directions

There are several directions which could lead to important findings on a practical level. Future research should investigate how my results from Study1, 2 and 3 generalize to other tasks and situations. Furthermore, it would be an important potential path to determine the personality and task features which lead to incorrect initial answers and to either one or the other type of bias. We have seen in Study 1 and Study 2 that some individuals gave biased answers because of the lack of intervening deliberate thinking while in other situations analytical thinking was the main cause of the bias. My final practical level remark is rooted in the findings of Study 4: it would be an essential next step to create a consensus- and intervention-based taxonomy of the choice architecture interventions which is integrated into the existing taxonomy of the greater field of behavioral interventions and to create a nudge specific reporting guideline to improve the quality of future choice architecture research.

Additionally, looking at my findings in the wide frame of the replication crisis, it can be concluded that the field of decision-making has a lot to do. We need to test and further refine previous models explaining human decision-making and explore factors influencing decisions. Furthermore, both the models and the interventions building on these models need to have a more nuanced understanding on the psychological processes and mechanisms leading to the decisions. Replications, refinements and process-level explorations are important for several reasons. They

lead to a more sophisticated and reliable model of human decision-making and, by that, let practitioners apply this scientific knowledge more effectively.

References

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology*, *27*(3), 379-387.

Aczel, B., Bago, B., Szollosi, A., Foldes, A., & Lukacs, B. (2015). Measuring Individual Differences in Decision Biases: Methodological Considerations. *Frontiers in Psychology, 6*:1770.

Aczel, B., Szollosi, A., & Bago, B. (2016). Lax monitoring versus logical intuition: The determinants of confidence in conjunction fallacy. *Thinking & Reasoning*, *22*(1), 99-117

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*(1), 124–140.

Arkes, H. R., Herren, L. T., & Isen, A. M. (1988). The role of potential loss in the influence of affect on risk-taking behavior. *Organizational behavior and human decision processes, 42*(2), 181-193.

Baca-Motes, K., Brown, A., Gneezy, A., Keenan, E. A., & Nelson, L. D. (2013). Commitment and behavior change: Evidence from the field. *Journal of Consumer Research*, *39*(5), 1070–1084.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Banks, A. P., & Hope, C. (2014). Heuristic and analytic processes in reasoning: An event-related potential study of belief bias. *Psychophysiology, 51,* 290–297.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioural and Brain Sciences, 30,* 241–256.

Baron, J. (1993). Why Teach Thinking?-An Essay. *Applied Psychology*, *42*(3), 191–214.

Baron, J. (2008). *Thinking and deciding*. New York: Cambridge University Press.

Baron, J., Badgio, P., & Gaskins, I. W. (1986). Cognitive style and its improvement: A normative approach. *Advances in the Psychology of Human Intelligence*, *3*, 173–220.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2014). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*. http://dx.doi.org/10.1016/j.jarmar.2014.09.003

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, *21*(4), 477–487.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., … Simel, D. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *Jama*, *276*(8), 637–639.

Behavioural Insights Ltd. (2015) *The Behavioural Insights Team Update Report 2013-15.* Retrieved from http://38r8om2xjhhl25mw24492dir.wpengine.netdna-cdn.com/wp-content/uploads/2015/08/BIT_Update-Report-Final-2013-2015.pdf

Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., & Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1810), 20150228.

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*(2), 186–196.

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, C. E. … Wollbrant. (2017). Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527-542.

Böckenholt, U. (2012). The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometrika*, *77*(2), 388–399.

Brandstätter, E., & Gussmack, M. (2013). The cognitive processes underlying risky choice. *Journal of Behavioral Decision Making*, *26*(2), 185–197.

Brosnan, M., Hollinworth, M., Antoniadou, K., & Lewton, M. (2014). Is Empathizing intuitive and Systemizing deliberative? *Personality and Individual Differences*, *66*, 39–43.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, *92*(5), 938-956.

Bucher, T., Collins, C., Rollo, M. E., McCaffrey, T. A., De Vlieger, N., Van der Bend, D., … Perez-Cueto, F. J. (2016). Nudging consumers towards healthier choices: a systematic review of positional influences on food choice. *British Journal of Nutrition*, *115*(12), 2252–2263.

Burk, D., Ingram, J. N., Franklin, D. W., Shadlen, M. N., & Wolpert, D. M. (2014). Motor effort alters changes of mind in sensorimotor decision making. *PLoS One*, *9*(3), e92681.

Campbell, N. C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., … Kinmonth, A. L. (2007). Designing and evaluating complex interventions to improve health care. *British Medical Journal*, *334*, 455–459.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*(3), 434–447.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making*, *5*(3), 182–191.

Carey, R.N., Evans, F., Horan, M., Johnston, M., West, R. & Michie, S. (2016). Describing the 'how' of behaviour change interventions: A taxonomy of modes of delivery. Paper presented at the UK Society for Behavioural Medicine Conference, Cardiff, UK.Retrieved from: https://www.ucl.ac.uk/human-behaviour-change/pdfs/tat-feedback.pdf

Caramazza, A., Anzellotti, S., Strnad, L., & Lingnau, A. (2014). Embodied cognition and mirror neurons: a critical assessment. *Annual review of neuroscience*, *37*, 1-15.

Chaiken, S. & Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. Guilford Press. New York.

Chan, A.-W., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., … Berlin, J. A. (2013). SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Annals of Internal Medicine*, *158*(3), 200–207.

Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P (2009) Reporting ofSample Size Calculation in Randomised Controlled Trials: Review. *British Medical Journal,* 338: b1732.

Chen, M. & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin, 25,* 215-224.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25–47.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, *4*(1), 20–33.

Cokely, E. T., Parpart, P., & Schooler, L. J. (2009). On the link between cognitive control and heuristic processes. In N. A. Taatgnen & H. Van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 2926–2931). Austen TX: The Cognitive Science Society.

Coles, M. G., Gratton, G., Bashore, T. R., Eriksen, C. W., & Donchin, E. (1985). A psychophysiological investigation of the continuous flow model of human information processing. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(5), 529–553.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal, 337*, 979–983.

Datta, S., & Mullainathan, S. (2014). Behavioral design: a new approach to development policy. *Review of Income and Wealth*, *60*(1), 7–35.

Davis, R., Campbell, R., Hildon, Z., Hobbs, L., & Michie, S. (2015). Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychology Review*, *9*(3), 323–344.

Del Missier, F., Mäntylä, T., & Bruin, W. B. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making*, *25*(4), 331–351.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, 7, 28–38.

De Neys, W., & Bonnefon, J.-F. (2013). The "whys" and "whens" of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*(4), 172–178.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299.

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 208–216.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? Perspectives on *Psychological Science, 6*(3), 274–290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology,* 5, 781.

Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., & Vlaev, I. (2012). Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, *33*(1), 264–277.

Dolan, P., Hallsworth, M., Halpern, D., King, D., & Vlaev, I. (2010). *Mindspace.* Institute for Government, the Cabinet Office. London: England, UK.

Durgin, F. H., Baird, J. A., Greenburg, M., Russell, R., Shaughnessy, K., & Waymouth, S. (2009). Who is being deceived? The experimental demands of wearing a backpack. *Psychonomic Bulletin & Review*, *16*(5), 964–969.

Ebeling, F., & Lotz, S. (2015). Domestic uptake of green energy promoted by opt-out tariffs. *Nature Climate Change,* 5, 868-871.

Elqayam, S., & Evans, J. S. B. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*(5), 233–248.

Epstein, S., & Pacini, R. (2001). The influence of visualization on intuitive and analytical information processing. *Imagination, Cognition and Personality, 20*(3), 195-216

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215-251.

Evans, J. St. B. T. (2009). How many dual process theories do we need: One, two, or many? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 33–54). City: Oxford University Press.

Evans, J. S. B. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology*, *87(2),* 223-240.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Evans, J. S. B. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339.

Evans, J. S. B. (2009). How many dual-process theories do we need? One, two, or many? In *In two minds: Dual processes and beyond* (pp. 33–54). New York, USA: Oxford University Press.

Evans, J. S. B., & Ball, L. J. (2010). Do people reason on the Wason selection task: A new look at the data of Ball et al. (2003*). Quarterly Journal of Experimental Psychology*, *63*(3), 434-441

Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, *11*(3), 295-306.

Evans, A. M., Dillon, K. D., Goldin, G., & Krueger, J. I. (2011). Trust and self-control: The moderating role of the default. *Judgment and Decision Making*, *6*(7), 697–705.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on psychological science, 8*(3), 223-241.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Fiedler, K. (2016). Functional research and cognitive-process research in behavioural science: An unequal but firmly connected pair. *International Journal of Psychology*, *51*(1), 64-71

Fiedler, K. (2017). What Constitutes Strong Psychological Science? The (Neglected) Role of Diagnosticity and A Priori Theorizing. *Perspectives on Psychological Science,12*(1), 46-61.

Fific, M., (2014) Double jeopardy in inferring cognitive processes. *Frontiers in Psychology*, http://doi.org/doi: 10.3389/fpsyg.2014.01130

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging*, *25*(2), 271-288.

Firestone, C. (2013). How "paternalistic" is spatial perception? Why wearing a heavy backpack doesn't—and couldn't—make hills look steeper. *Perspectives on Psychological Science*, *8*(4), 455–473.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316-344.

Förster, J. & Strack, F. (1998). Motor actions in retrieval of valenced information: II. Boundary conditions for motor congruence effects. *Perceptual and Motor Skills, 86,* 1423-1426.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *The Journal of Economic Perspectives*, *19*(4), 25–42.

Freeman, J. B. (2014). Abrupt category shifts during real-time person perception. *Psychonomic Bulletin & Review*, *21*(1), 85-92.

Freeman, J., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59.

Gaker, D., Zheng, Y., & Walker, J. (2010). Experimental economics in transportation: Focus on social influences and provision of information. *Transportation Research Record: Journal of the Transportation Research Board*, (2156), 47–55.

Gardner, M., & Steinberg, L. (2005). Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: an experimental study. *Developmental psychology, 41*(4), 625-635.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*(6080), 493–496.

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making*, *9*(1), 15–34.

Gigerenzer, G. (2011). Personal reflections on theory and psychology. Theory & Psychology, 20, 733–743.

Glanz, K., & Bishop, D. B. (2010). The role of behavioral science theory in development and implementation of public health interventions. *Annual Review of Public Health*, *31*, 399–418.

Gonzales, J. E., & Cunningham, C. A. (2015). *The promise of pre-registration in psychological research.* Retrieved from the website ofPsychological Science: http://www.apa.org/science/about/psa/2015/08/pre-registration.aspx.

Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the revolution to embodiment 25 years of cognitive psychology. *Perspectives on Psychological Science*, *8*(5), 573–585.

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87,* 11–22.

Goldinger, S. D., Papesh, M. H., Barnhart, A. S., Hansen, W. A., & Hout, M. C. (2016). The poverty of embodied cognition. *Psychonomic Bulletin & Review*, *23*(4), 959–978.

Grant, S. P., Mayo-Wilson, E., Melendez-Torres, G. J., & Montgomery, P. (2013). Reporting quality of social and psychological intervention trials: a systematic review of reporting guidelines and trial publications. *PloS one, 8*(5), e65442.

Gürçay, B., & Baron, J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning, 23*(1), 49-80.

Hansen, P. G., Skov, L. R., & Skov, K. L. (2016). Making healthy choices easier: regulation versus nudging. *Annual Review of Public Health*, *37*, 237–251.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, *8*(3), 188–201.

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, *18*(3), 384–401. https://doi.org/10.1177/1368430214538325

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., … Johnston, M. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal, 348*, doi:10.1136/bmj.g1687

Hollands, G. J., Shemilt, I., Marteau, T. M., Jebb, S. A., Kelly, M. P., Nakamura, R., … Ogilvie, D. (2013). Altering choice architecture to change population health behaviour: a large-scale conceptual and empirical scoping review of interventions within micro-environments. *BMC Public Health*, *13*, 1218. http://dx.doi.org/10.1186/1471- 2458-13-1218.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory.* Appleton-Century Company, Incorporated. Oxford, UK.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124

Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection versus late correction. In. S. Chaiken & Y. Trope (Eds), *Dual-process theories in social psychology* (pp. 383-400). New York: Guilford.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, *12*(5), 852–857.

Jachimowicz, J. M., Duncan, S., Weber, E. U., Johnson, E. J., (2017) *When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects.* Manuscript submitted for publication.

Jeffreys, H. (1961). *The Theory of Probability.* Oxford, UK: Oxford University Press.

Johnson E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302,* 1338–1339.

Johnson, E. J., Shu, S. B., Dellaert, B. G., Fox, C., Goldstein, D. G., Häubl, G., … Schkade, D. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, *23*(2), 487–504.

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64.

Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs: General and Applied*, *78*(1), 1-37.

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. Perspectives on psychological science, 4, 533–550.

Kiani, R., Cueva, C. J., Reppas, J. B., & Newsome, W. T. (2014). Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Current Biology*, *24*(13), 1542–1547.

Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods, 49*(5), 1652-1667. https://doi.org/10.3758/s13428-017-0900-z

Kieslich, P. J., Wulff, D. U., Henninger, F., Haslbeck, J. M. B., & Schulte-Mecklenbeck, M. (2016). Mousetrap: An R package for processing and analyzing mouse-tracking data. https://doi.org/10.5281/zenodo.596640

Kokis, J. V., Macpherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, *83*(1), 26–52.

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making, 8*(5), 527-539.

Koop, G. J., & Johnson, J. G. (2011). Response dynamics: A new window on the decision process. *Judgment and Decision Making, 6*(8), 750-758.

Koop, G. J., & Johnson, J. G. (2013). The response dynamics of preferential choice. *Cognitive psychology, 67*(4), 151-185.

Kovacs, K., & Temesvári, E. (2016). Számítógépes, adaptív IQ-mérés: Egy gyakorlati példa. [Computerized, adaptive IQ measurement: an example from Hungary]. *Magyar Pszichológiai Szemle*, *71*(1), 143–163.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1 - 26.

Låg, T., Bauger, L., Lindberg, M., & Friborg, O. (2014). The Role of Numeracy and Intelligence in Health-Risk Estimation and Medical Data Interpretation. *Journal of Behavioral Decision Making*, *27*(2), 95–108.

Lenzer, J., Hoffman, J. R., Furberg, C. D. & Ioannidis, J. P. A. (2013) Ensuring the integrity of clinical practice guidelines: a tool for protecting patients. *British Medical Journal, 347,* f5535.

Lepora, N. F., & Pezzulo, G. (2015). Embodied choice: how action influences perceptual decision making. *PLoS Comput Biol*, *11*(4), e1004110.

Li, M., & Chapman, G. B. (2013). Nudge to health: Harnessing decision research to promote health behavior. *Social and Personality Psychology Compass*, *7*(3), 187–198.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*(4), 361–381.

Lindsay, S. D. (2016). *Preregistration of Research Plans*. Retrieved from the website of Psychological Science: http://www.apa.org/science/about/psa/2015/08/pre-registration.aspx.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135.

Mahon, B. Z. (2015). What is embodied about cognition? *Language, Cognition and Neuroscience*, *30*(4), 420–429.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, *102*(1), 59–70.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology*, *105*(3), 353-355.

Mata, A., Schubert, A.-L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition*, *133*(2), 457–463.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324.

Mathur, M. B., Bart-Plange, D., Aczel, B., Bernstein, M., Ciunci, A., Ebersole, C., … Falcao, F. (2017). *Registered multi-site replication of tempting-fate effects in Risen & Gilovich (2008).* Manuscript submitted for publication.

McLeroy, K. R., Garney, W., Mayo-Wilson, E., & Grant, S. (2016). Scientific reporting: Raising the standards. *Health Education & Behavior, 43*(5), 501-508.

Meyer, A., Spunt, R., & Frederick, S. (2015). *The bat and ball problem.* Manuscript in preparation.

Michie, S., Carey, R. N., Johnston, M., Rothman, A. J., de Bruin, M., Kelly, M. P., & Connell, L. E. (2016). From theory-inspired to theory-based interventions: A protocol for developing and testing a methodology for linking behaviour change techniques to theoretical mechanisms of action. *Annals of Behavioral Medicine.* doi:10.1007/s12160-016-9816-6

Michie, S., & Johnston, M. (2012). Theories and techniques of behaviour change: Developing a cumulative science of behaviour change. *Health Psychology Review*, *6*(1), 1–6.

Michie, S., Johnston, M., Francis, J., Hardeman, W., & Eccles, M. (2008). From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. *Applied Psychology*, *57*(4), 660–680.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., … Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, *46*(1), 81–95.

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *Lancet, 357* (2001), 1191–1194.

Moher, D., Schulz, K., Simera, I., & Altman, D. (2010). Guidance for developers of health research reporting guidelines. *PLoS Medicine*, 7, 2.

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6-18.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). Package 'BayesFactor'. R package, version 0.9.12-2. https://cran.r-project.org/web/packages/BayesFactor/index.html

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour,* 1, 0021.

Münscher, R., Vetter, M., & Scheuerle, T. (2015). A review and taxonomy of choice architecture techniques. *Journal of Behavioral Decision Making*. http://doi.org/10.1002/bdm.1897

Nettle, D., Nott, K., & Bateson, M. (2012). "cycle thieves, we are watching you": impact of a simple signage intervention against bicycle theft. *PloS One*, *7*(12), 51738.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104). Prentice-Hall. New Jersey, US.

Nosek, B. A. & Lakens (2014). D. Registered Reports: a method to increase the credibility of published results. *Social Psychology. 45,* 137–141.

Nørnberg, T. R., Houlby, L., Skov, L. R., & Peréz-Cueto, F. J. A. (2015). Choice architecture interventions for increased vegetable intake and behaviour change in a school setting: a systematic review. *Perspectives in Public Health*, *136*(3), 132–142. http://doi.org/10.1177/1757913915596017

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943.

Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. *BMJ*, *332*(7538), 413–416.

Pachur, T., & Spaar, M. (2015). Domain-specific preferences for intuition and deliberation in decision making. *Journal of Applied Research in Memory and Cognition*, *4*(3), 303–311.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*(6), 972-987.

Painter, J. E., Borba, C. P., Hynes, M., Mays, D., & Glanz, K. (2008). The use of theory in health behavior research from 2000 to 2005: a systematic review. *Annals of Behavioral Medicine, 35*(3), 358.

Palfi, B., Kieslich, P. J., Szaszi, B., Wulff, D. U., & Aczel, B. (2017) *Development and empirical comparison of mouse-tracking methods assessing changes of mind in reasoning and decision making*. Manuscript in preparation.

Papesh, M. H. (2015). Just out of reach: On the reliability of the action-sentence compatibility effect. *Journal of Experimental Psychology: General*, *144*(6), e116.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making*, *18*(1), 1–27.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163–177.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2015). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, http://doi.org/doi: 10.3758/s13428-015-0576-1

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.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition, 124,* 101–106.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, *24*(6), 425–432.

Pennycook, G., & Ross, M. R. (2016). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, http://doi.org/doi: 10.3389/fpsyg.2016.00009

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, *19*(3), 528–534.

Perkins, D. (1995). *Outsmarting IQ: The emerging science of learnable intelligence*. New York: Free Press.

Peters, E. (2012). Beyond comprehension the role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31–35.

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., & Mertz, C. K. (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review*, *64*(2), 169–190.

Piazza, J., & Sousa, P. (2013). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science*, 1948550613492826.

Pinto-Prades, J. L., Martinez-Perez, J. E., & Abellán-Perpiñán, J. M. (2006). The Influence of the Denominator neglect Phenomenon on the Elicitation of Health States Utilities. *Judgment and Decision Making, 1*(2), 118-133.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2015). The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). *Journal of Behavioral Decision Making*, http://doi.org/10.1002/bdm.1883

R Core Team (2016). R*: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reisen, N., Hoffrage, U., & Mast, F. W. (2008). Identifying decision strategies in a consumer choice situation. *Judgment and Decision Making*, *3*(8), 641–658.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, *135*(6), 943-973.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641.

Rozin, P., Scott, S., Dingley, M., Urbanek, J. K., Jiang, H., & Kaltenbach, M. (2011). Nudge to nobesity I: Minor changes in accessibility decrease food intake. *Judgment and Decision Making*, *6*(4), 323-332.

Samson, A. (Ed.). (2016). *The Behavioral Economics Guide 2016 (with an introduction by Gerd Gigerenzer)*. Retrieved from http://www.behavioraleconomics.com

Scherbaum, S., & Kieslich, P. J. (in press). Stuck at the starting line: How the starting procedure influences mouse-tracking data. *Behavior Research Methods*.

Schubert, T. W. (2004). The power in your hand: Gender differences in bodily feedback from making a fist. *Personality and Social Psychology Bulletin, 30,* 757-769.

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., & Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science*, *26*(5), 442-450.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *British Medical Journal, 340* (2010), pp. 698–702.

Shiffrin, R. M., and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, **84**, 127–90.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129-138.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359-1366.

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology, 6:*532, http://doi.org/doi: 10.3389/fpsyg.2015.00532

Singmann, H., Bolker, B., & Westfall, J. (2015). Afex: analysis of factorial experiments. R package version 0.13–145. https://cran.r-project.org/web/packages/afex/index.html

Skov, L. R., Lourenco, S., Hansen, G. L., Mikkelsen, B. E., & Schofield, C. (2013). Choice architecture as a means to change eating behaviour in self-service settings: a systematic review. *Obesity Reviews*, *14*(3), 187–196.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin. 119*. 3-22.

Smith, N. C., Goldstein, D. G., & Johnson, E. J. (2013). Choice without awareness: Ethical and policy implications of defaults. *Journal of Public Policy & Marketing*, *32*(2), 159–172.

Social and Behavioral Science Team. (2015). *Social and Behavioral Sciences Team 2015 Annual Report*. Retrieved from: https://sbst.gov/assets/files/2015-annual-report.pdf

Soll, J. B., Milkman, K. L., & Payne, J. W. (2016). A Users' guide to debiasing. In G.B. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgement and decision making* (pp. 975-985). John Wiley & Sons Ltd.

Sousa Lourenço, J., Ciriolo, E., Rafael Almeida, S., & Troussard, X. (2016). *Behavioural insights applied to policy: European Report 2016*.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398.

Stupple, E. J., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning*, *19*(1), 54–77.

Stanovich K. E. (2009). Distinguishing the reflective, algorithmic and autonomous minds: Is it time for a tri-process theory? In Evans J. St. B. T., Frankish K. (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford, England: Oxford University Press.

Stanovich, K. E., & West, R. F. (1998a). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161-188.

Stanovich, K. E., & West, R. F. (1998b). Who uses base rates and P (D/~ H)? An analysis of individual differences. *Memory & Cognition*, *26*(1), 161–179.

Stanovich, K. E., and West, R. F. (1998c). Cognitive ability and variation in selection task performance. *Thinking & Reasoning*, **4**, 193–230.

Stanovich, K. E., and West, R. F. (1998d). Individual differences in framing and conjunction effects. *Thinking & Reasoning*, **4**, 289–317.

Stanovich, K. E., & West, R. F. (2001). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(05), 645–665.

Stanovich, K. E., and West, R. F. (2007). Natural my side bias is independent of cognitive ability. *Thinking & Reasoning*, **13**, 225–47.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672-695.

Stanovich, K. E., Toplak, M. E., & West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. *Advances in Child Development and Behavior*, *36*, 251-285.

Strack, F. Martin, L.L. & Schwarz, N. (1988). Priming and communication: The social determinants of information use in judgments of life-satisfaction. *European Journal of Social Psychology*. 18. 429-442.

Stupple, E. J., Gale, M., & Richmond, C. R. (2013). Working Memory, Cognitive Miserliness and Logic as Predictors of Performance on the Cognitive Reflection Test. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1396 –1401). Austin, TX: Cognitive Science Society.

Svedholm-Häkkinen, A. M. (2015). Highly reflective reasoners show no signs of belief inhibition. *Acta Psychologica*, *154*, 69–76.

Szaszi, B. (2016) The role of expertise and preference behind individuals' tendency to use intuitive decision style. *Journal of Applied Research in Memory and Cognition.* http://dx.doi.org/10.1016/j.jarmac.2016.06.006

Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A., & Aczel, B. (2016). *Annotated bibliography of the Choice Architecture interventions between 2008 to 2016*. Manuscript.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: exploring the ways individuals solve the test. *Thinking & Reasoning*, *23*(3), 1–28.

Szollosi, A., Bago, B., Szaszi, B., Aczel, B. (manuscript in preparation) Verification in Reasoning: Top-Down Influences on Confidence on the Bat-and-Ball Problem.

Szollosi, A., Bago, B., Szaszi, B., & Aczel, B. (2017). Exploring the Determinants of Confidence in the Bat-and-Ball Problem. *Acta Psychologica, 180, 1-7.*

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2014). *Choice architecture*. (E. Shafir, Ed.). Available at http://ssrn.com/abstract=2536504

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. Evans & K. Frankish (Eds*.), In two minds: Dual processes and beyond.* Oxford: Oxford University Press.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244.

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. G. (2017). Do Smart People Have Better Intuitions? *Unpublished manuscript.*

Thompson, V. A., Turner, J. P., & Pennycook, G. (2011). Intuition, reason and metacognition. *Cognitive Psychology*, *63(3)*, 107–140

Thompson, V. A., Turner, J. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128 (2)*, 237–251.

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

Toplak, M. E. and Stanovich, K. E. (2003). Associations between my side bias on an informal reasoning task and amount of post-secondary education. *Applied Cognitive Psychology*, **17**, 851–60.

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.

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.

Tor, A., & Bazerman, M. H. (2003). Focusing failures in competitive environments: Explaining decision errors in the Monty Hall game, the acquiring a company problem, and multiparty ultimatums. *Journal of Behavioral Decision Making*, *16*(5), 353–374.

Travers, E., Rolison, J. J., & Feeney, A. (2016). *The Time Course of Conflict on the Cognitive Reflection Test. Cognition*, *150,* 109–18.

Tversky, A. & Kahneman, D. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*. 5(2). 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van Prooijen, F.-W., Karremans, J. C., & van Beest, I. (2006). Procedural justice and the hedonic principle: How approach versus avoidance motivation influences the psychology of voice. *Journal of Personality and Social Psychology, 91,* 686-697.

Vlaev, I., & Dolan, P. (2015). Action change theory: A reinforcement learning perspective on behaviour change. *Review of General Psychology, 19,* 69-95.

Vlaev, I., King, D., Dolan, P., & Darzi, A. (2016). The Theory and Practice of "Nudging": Changing Health Behaviors. *Public Administration Review*, 76(4), 550-561.

Wason, P. C., & Evans, J. S. B. (1975). Dual processes in reasoning?. *Cognition, 3*(2), 141-154.

Webb, T., Joseph, J., Yardley, L., & Michie, S. (2010). Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *Journal of medical Internet research, 12*(1)*,* e4

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*(2), 198–212.

Welsh, M., Burns, N., & Delfabbro, P. (2013). The Cognitive Reflection Test: how much more than numerical ability. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1396 –1401). Austin, TX: Cognitive Science Society.

West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*(2), 243–251.

Wilson, A. L., Buckley, E., Buckley, J. D., & Bogomolova, S. (2016). Nudging healthier food and beverage choices through salience and priming. Evidence from a systematic review. *Food Quality and Preference*, *51*, 47–64.

Yamagishi, K. (1997). When a 12.86% mortality is more dangerous than 24.14%: Implications for Risk Communication. *Applied Cognitive Psychology, 11,* 495-506.

Zipf, G. K. (1950). Human behavior and the principle of least effort: An introduction to human ecology. *Language*, *26*(3), 394–401.