Activating reflective thinking with decision justification and debiasing training

Ozan Isler*   Onurcan Yilmaz[†]   Burak Dogruyol[‡]

Abstract

Manipulations for activating reflective thinking, although regularly used in the literature, have not previously been systematically compared. There are growing concerns about the effectiveness of these methods as well as increasing demand for them. Here, we study five promising reflection manipulations using an objective performance measure — the Cognitive Reflection Test 2 (CRT-2). In our large-scale preregistered online experiment (N = 1,748), we compared a passive and an active control condition with time delay, memory recall, decision justification, debiasing training, and combination of debiasing training and decision justification. We found no evidence that online versions of the two regularly used reflection conditions — time delay and memory recall — improve cognitive performance. Instead, our study isolated two less familiar methods that can effectively and rapidly activate reflective thinking: (1) a brief debiasing training, designed to avoid common cognitive biases and increase reflection, and (2) simply asking participants to justify their decisions.

Keywords: cognitive reflection, time delay, memory recall, decision justification, debiasing training

1 Introduction

The distinction between reflective and intuitive thinking guides a wide range of research questions in modern behavioral sciences. The dual-process model of the mind provides the leading theoretical framework for these questions by positing that cognition is based on two fundamentally distinct types of processes (Evans & Stanovich, 2013; Morewedge & Kahneman, 2010). Type 1 processes include the automatic, effortless, and intuitive thinking that we share with our evolutionary ancestors, whereas Type 2 processes include the controlled, effortful, and reflective thinking specific to humans (Kahneman, 2011). Although the assumption of the dual-process model that the two cognitive processes are independent has recently come under scrutiny (Baron, Scott, Fincher & Metz, 2015; Białek & De Neys, 2016; Klein, 2011; Pennycook, Fugelsang & Koehler, 2015; Thompson, Evans & Frankish, 2009; Trémolière & Bonnefon, 2014), it is well-established that the relative extent of reflection vs. intuition constituting a decision-making process can nevertheless strongly influence beliefs and behaviors (e.g., ideological, re-

ligious, and conspirational beliefs, and economic, moral, and health behaviors; Gervais et al., 2018; Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2013; Pennycook, Cheyne, Seli, Koehler & Fugelsang, 2012; Rand, 2016; Swami, Voracek, Stieger, Tran & Furnham, 2014; Yilmaz & Isler, 2019; Yilmaz & Saribay, 2017a, 2017b).

Surprisingly, the relative effectiveness of reflection and intuition manipulations used in behavioral research remains largely unknown (Horstmann, Hausmann & Ryf, 2009; Myrseth & Wollbrant, 2017). We are aware of only one (unpublished) experimental comparison of intuition manipulations in cognitive performance (Deck, Jahedi & Sheremeta, 2017), and no previous experimental study that has systematically compared alternative reflection manipulations. The presumed effectiveness of reflection manipulations used in the literature can be questioned since baseline cognitive functions tend to be intuitive and motivating people to pursue an effortful activity such as reflection can be difficult (e.g., Kahneman, 2011). Here, we provide possibly the first systematic methodological comparison of regularly used and promising reflection manipulations.

Another reason for the missing methodological evidence is the frequent lack of control conditions, which stems from a reliance on experimental comparisons of intuition and reflection manipulations as the basis for hypothesis testing. Without these controls, the question of whether experimental results are due to activation of intuitive or reflective processes cannot be answered (e.g., Isler, Maule & Starmer, 2018; Rand, 2016). Similarly, studies that rely on the two-response paradigm, where an initial (relatively more intuitive) response is elicited before a second (relatively less intuitive and more reflected) response, often lack a control

---

*Centre for Behavioral Economics, Society and Technology, School of Economics and Finance, Queensland University of Technology, Brisbane, Australia. Email: ozan.isler@qut.edu.au. ORCID: 0000-0002-4638-2230.

†Department of Psychology, Kadir Has University, Istanbul, Turkey. ORCID: 0000-0002-6094-7162.

‡Department of Psychology, Altinbas University, Istanbul, Turkey. ORCID: 0000-0002-3469-590X.

condition (e.g., Bago & De Neys, 2017). As a recent exception, Lawson, Larrick, and Soll (2020) employ slow and fast thinking prompts (without time-limits) and find that slow thinking has limited positive effect on cognitive performance compared to a control condition. Given its importance, we also employ control conditions in the current study.

Studies using intuition and reflection manipulations often do not directly test whether cognitive processes were activated in the intended directions. While some have checked the direct effects of their manipulations on cognitive performance (e.g., Deppe et al., 2015; Lawson et al., 2020; Yilmaz & Saribay, 2016), subjective self-report questions and behavioral measures such as response times are frequently relied on as alternative manipulation checks (Rand, Greene & Nowak, 2012; Yilmaz & Isler, 2019). The lack of performance measures would be misleading if, rather than thinking reflectively about the problem at hand, participants were to rely on their own lay theories about reflection (Saribay, Yilmaz & Körpe, 2020) or if they were to respond in socially desirable ways (Grimm, 2010). Consistent with the existence of such methodological problems, Saribay et al. (2020) found intuition and reflection primes to affect self-reported thinking style but not actual performance in the commonly used Cognitive Reflection Test (CRT, Frederick, 2005). Even the regularly used objective performance measures — such as when differences in response times are used to check whether time-limit manipulations have impacted behavior (e.g., Isler et al., 2018; Rand et al., 2012) — may not always provide direct and convincing evidence about whether and how cognitive processes have been manipulated (Krajbich, Bartling, Hare & Fehr, 2015).

Therefore, the effect of reflection manipulations should be observed on well-established measures of cognitive performance — such as the CRT (Frederick, 2005) and the CRT-2 (Thomson & Oppenheimer, 2016). Providing evidence of their ability to predict the domain-general features of reflection, test scores on these two tasks have been shown to correlate with a wide-range of cognitive performance measures in the lab (e.g., syllogistic reasoning and heuristics-and-biases problems) and in the field (e.g., standardized academic test scores and university course grades) (Lawson et al., 2020; Meyer, Zhou & Shane, 2018; Thomson & Oppenheimer, 2016; Toplak, West, & Stanovich, 2011). Numerous other widely-used reasoning problems, such as the conjunction fallacy (Tversky & Kahneman, 1983), probability matching (Stanovich & West, 2008) and base rate neglect (Kahneman & Tversky, 1973), can also be used to measure the effects of manipulations on cognitive performance (e.g., Lawson et al., 2020). Among these alternatives, we chose CRT-2 as our performance measure because participants are less likely to be familiar with it, thereby minimizing problems such as ceiling effects, and because its reliance on numeracy skills is less than that of CRT, which can confound the interpretation of scores (see discussion in Thomson & Oppenheimer, 2016).

Despite these advantages, the CRT-2 arguably captures only some of the specific features of cognitive reflection directly, such as attention to detail and careful reading. Hence, the immediate effects of the reflection manipulations found in our study can be limited to these features of reflection, as we further detail in the Discussion.

The increased reliance on online experiments provides another reason to study the effectiveness of reflection manipulations, namely, to test their robustness in this novel research environment. Online labor markets such as Amazon Mechanical Turk as well as professionally maintained research participant pools such as Prolific have been shown to provide internally valid experimental tests in settings less artificial and more anonymous than the laboratory (Horton, Rand & Zeckhauser, 2011; Palan & Schitter, 2018; Peer, Brandimarte, Samat & Acquisti, 2017), but online experiments can also suffer from idiosyncratic drawbacks such as non-compliance with treatments and asymmetry in dropout rates (Arechar, Gächter & Molleman, 2018; Isler et al., 2018). These problems may be more acute for cognitively demanding tasks such as the reflection manipulations that we study here, especially in online decision environments that can be distracting to participants (Dandurand, Shultz & Onishi, 2008). For example, providing participants with monetary incentives has been shown to result in high rates of compliance with time-limits (Isler et al., 2018) and reflective thinking (Lawson et al., 2020) in online experiments. With these considerations in mind, we compare five tasks that are simple and fast enough to be used in online experiments, and we use monetary incentives to motivate compliance for the task instructions.

Numerous experimental tasks for promoting reflective thinking are currently in use. Some of these tasks, introduced in once-acceptable small-sample studies, are now known to be unreliable. For example, the perceptual disfluency method (e.g., the use of hard-to-read-fonts to promote reflection), the scrambled sentence task that primes participants with words such as "reason" and "rational", and the task that aims to prime reflection by showing participants a picture of Rodin's *The Thinker* (Gervais & Norenzayan, 2012; Song & Schwarz, 2008) all failed to manipulate reflective thinking in recent large-sample replication attempts (Bakhti, 2018; Deppe et al., 2015; Meyer et al., 2015; Sanchez, Sundermeier, Gray & Calin-Jageman, 2017; Sirota, Theodoropoulou & Juanchich, 2020). In addition, researchers sometimes attempt to activate reflective thinking by having participants complete tasks (e.g., the CRT) that are originally designed to measure thinking style, but the effects of such unestablished approaches tend to be unreliable too (Yonker, Edman, Cresswell & Barrett, 2016). Instead, to make the most use of our experimental resources, we here focus on methods that are specifically designed to manipulate reflection and that are not known to be unreliable.

One of the most frequently used reflection manipulations is to put time-limits on decision-making processes (Horstmann, Ahlgrimm & Glöckner, 2009; Maule, Hockey & Bdzola, 2000; Spiliopoulos & Ortmann, 2018). In this method, participants in a time pressure condition, prompted to decide within a time-limit (e.g., 10 seconds), are compared to those in a time delay condition, who are either asked to think or forced to wait for a certain duration (e.g., 20 seconds) before submitting decisions (Capraro, Schulz & Rand, 2019; Rand, 2016; Suter & Hertwig, 2011). Although the time delay condition is assumed to induce reflective answers relative to the time pressure condition, the usual lack of a control condition without time-limits prohibits the identification of whether it is time pressure or time delay that affects decision-making. Only a few studies have used control conditions to isolate the influence of time delay (e.g., Everett, Ingbretsen, Cushman & Cikara, 2017). Nevertheless, the exact effect of time delay arguably remains unclear even with a control condition, as it may be difficult to distinguish between increased reliance on reflective processes and dilution of emotional responses (Neo, Yu, Weber & Gonzalez, 2013; Wang et al., 2011). Given its prominence as the most frequently used cognitive process manipulation, we here use time delay as one of our experimental conditions, and we also explore the role of emotional responses.

Another frequently used technique for activating reflection is memory recall (Cappelen, Sørensen & Tungodden, 2013; Forstmann & Burgmer, 2015; Ma, Liu, Rand, Heatherton & Han, 2015; Rand et al., 2012; Shenhav, Rand & Greene, 2012). In this method, participants are usually asked to write a paragraph describing a personal experience where reliance on careful reasoning led to a good outcome, with the expectation that the explicit priming of these memories would motivate reflection. Although a recent high-powered study failed to find an effect of this priming method on a cognitive performance measure (Saribay et al., 2020), this null result may have been a result of the low rates of compliance with the task instructions (see Shenhav et al., 2012). Similar difficulties in achieving high rates of compliance have been observed when using time-limits to activate reflection (Tinghog et al., 2013), and monetary incentives have successfully been implemented to resolve this problem (Isler et al., 2018; Kocher & Sutter, 2006). Building on these findings, we adapt this task to the online context and, as with other tasks tested in the study, use monetary incentives to motivate compliance.

In the third reflection manipulation that we test here, we simply ask participants to justify their answers by writing an explanation of their reasoning. Across multiple studies employing the classic Asian disease problem (Miller & Fagley, 1991; Sieck & Yates, 1997; Takemura, 1994), the decision justification task has been found to reduce framing effects effectively. Asking for justification or elaboration was found to be even more effective than monetary incentives (Vieider,

2011), and its effectiveness has been validated across multiple decision-making contexts, including health (Almashat, Ayotte, Edelstein & Margrett, 2008) and consumer choice (Cheng, Wu & Lin, 2014). Justification prompts can motivate reflection by generating feelings of higher levels of responsibility for one's decisions as well as expectations of their scrutiny by others. However, the effectiveness of the justification task has been questioned (Belardinelli, Bellé, Sicilia & Steccolini, 2018; Leboeuf & Shafir, 2003). Additional findings have suggested that the effectiveness of decision justification is task-dependent (Leisti, Radun, Virtanen, Nyman, & Häkkinen, 2014) and that it may even harm decisions (Igou & Bless, 2007), especially in specific contexts prone to motivated reasoning (Christensen, 2018; Sieck, Quinn & Schooler, 1999). Given the promising but mixed findings on the effectiveness of the justification task, we used this simple technique as an alternative reflection manipulation.

For the fourth reflection task tested here, we develop a novel training procedure for the online context consistent with well-established debiasing principles (Lewandowsky, Ecker, Seifert, Schwarz & Cook, 2012). We modify a debiasing training task that was previously tested in the laboratory with promising results (Yilmaz & Saribay, 2017a, 2017b). The lab version of the task provides participants with a 10-minute training on noticing and correcting cognitive biases: it first elicits the Cognitive Reflection Test (Frederick, 2005) and various base-rate problems (De Neys & Glumicic, 2008) and then provides feedback on the correct answers and their explanations (also see Morewedge et al., 2015; Stephens, Dunn, Hayes & Kalish, 2020). While previous studies using debiasing training have been successful (Sellier, Scopelliti & Morewedge, 2019), its lengthy and complicated exercises have so far precluded its systematic use in online experiments.

In short, alternative reflection manipulations have not yet been experimentally compared using an actual performance measure and behavioral research methods lack reliable reflection manipulations that can be used in online experiments. Here, we use CRT-2 scores as the cognitive performance measure and compare the effects of five promising manipulations on reflective thinking in a high-powered between-subjects experiment. The five reflection manipulations include the time delay condition (R1), the memory recall task (R2), the decision justification task (R3), and the debiasing training (R4) described above as well as a combined task that includes both the debiasing training and the decision justification tasks (R5). We compare these five reflection conditions with two control groups: the passive control condition (C1) where participants received no treatment prior to taking part in CRT-2, and the active control condition (C2) where participants were assigned neutral reading and writing tasks to provide comparability with the reflection conditions.

TABLE 1: Overview of reflection manipulations.

Manipulation	Task description	Completed (as % of recruited)
Passive control (C1):	No manipulations or active controls	262 (99%)
Active control (C2):	Neutral reading and writing task	255 (96%)
Time delay (R1):	Thinking carefully for at least 20 seconds for each question	262 (99%)
Memory recall (R2):	Describing a time when reflection was beneficial	210 (79%)
Decision justification (R3):	Justifying answers to each question	256 (97%)
Debiasing training (R4):	Learning about and describing three common cognitive biases	252 (95%)
R3 + R4 (R5):	Combination of debiasing training and justification	251 (95%)

Using this experimental setup, we test three preregistered hypotheses on the effect of manipulations on reflective thinking as measured by the CRT-2 scores. First, we predicted that the CRT-2 scores in the five reflection conditions (R1 to R5) will be higher than the two control conditions (C1 to C2). Second, we predicted that the CRT-2 scores in conditions with debiasing training (R4 and R5) will be higher than the reflection conditions without debiasing training (R1, R2 and R3) because they are based on proven debiasing techniques, including repeated explanations of cognitive biases and warnings against potential future mistakes (Lewandowsky et al., 2012). Third, we expected that the combination of debiasing training and decision justification manipulations can motivate even higher reflection by prompting participants to apply debiasing techniques when providing justifications for their decisions on the CRT-2 items. Accordingly, we predicted that the CRT-2 scores in the debiasing training condition with justification (R5) will be higher than the debiasing training condition without justification (R4).

In addition to testing these hypotheses, we report various exploratory analyses. We investigate response times and study the role of task compliance in driving the treatment effects. We then contrast CRT-2 scores with self-report measures of reflection. We conjectured that a discrepancy between these two measures, where self-reported reflection is not supported by actual performance, could indicate socially desirable responding. There is limited but suggestive evidence that reflection manipulations such as time limits can influence affect (Isler et al., 2018; Maule et al., 2000). Therefore, we also explore whether the effects of treatments on cognitive performance align with differences in effects on emotional responses.

2   Method

Using a between-subjects design, we experimentally compared five reflection manipulations and two control conditions. Participants were blind to the experimental conditions, and each participant was randomly assigned to one of seven conditions (see Table 1). The experiment was preregistered at the Open Science Framework (OSF) (https://osf.io/6axuz). The experimental materials, the dataset, and the analysis code are available at the OSF study site (https://osf.io/k495r/).

2.1   Participants

Participants were recruited online via Prolific (http://www.prolific.co/, Palan & Schitter, 2018) and recruitment was restricted to fluent English-speaking UK residents who were 18 or older. As preregistered, participants with incomplete data were excluded from the dataset prior to analysis (n = 107). None of the excluded participants had completed the CRT-2. Hence, their inclusion in the analysis does not change the results. We analyze data from 1,748 unique participants with complete submissions (M_{age} = 33.58, SD_{age} = 11.50; 71.1% female). In addition to a participation fee of £0.40, participants were paid £0.20 for compliance with task instructions.

2.2   Planned sample size

We planned for a powerful test (1-β = 0.90) to identify small effects of manipulations (f = 0.10) in a one-way ANOVA model with seven conditions and standard Type I error rate (α = 0.05). Using G*Power 3.1.9.2 (Faul, Erdfelder, Buchner & Lang, 2009), we estimated our target sample size to include at least 1750 complete submissions.

2.3   Procedure

To increase compliance with the experimental tasks, participants were informed that they would earn an additional £0.20 if they closely followed the task instructions. Five of the seven conditions were designed to activate cognitive reflection (R1 to R5), whereas the other two conditions were designed as controls (C1 and C2). In all conditions, participants completed the Cognitive Reflection Test (CRT-2;

Thomson & Oppenheimer, 2016), which provides a less familiar and less numerical alternative to the original CRT (Frederick, 2005). CRT-2 includes four questions that are designed to trigger a spontaneous but incorrect response and reliance on cognitive reflection is operationalized as resistance to this initial response (e.g., "If you're running a race and you pass the person in second place, what place are you in?"). Hence, individual CRT-2 scores range from 0 to 4. Cronbach's α for the four CRT-2 items was .54, in line with the original CRT (Baron et al., 2015). As we next describe in detail, the reflection manipulations were implemented during the CRT-2 for R1 and R3 and before the CRT-2 for R2 and R4, whereas participants in R5 were exposed to reflection manipulations both before and during the CRT-2.

In the first reflection manipulation (R1), the time delay condition, participants were asked to think for at least 20 seconds before answering each CRT-2 question. Each question screen displayed a reflection prompt ("Carefully consider your answer") and a timer counting up from zero seconds. Consistent with its regular use (Bouwmeester et al., 2017; Isler et al., 2018; Rand, 2016; Rand et al., 2012), it was technically possible to submit answers within 20 seconds, which allows checking that time delay instructions motivate behavior change (Horstmann, Hausmann, et al., 2009). The average rate of compliance with time-limits across the four questions was 67%.

The second reflection condition (R2), the memory recall task, was based on Shenhav et al. (2012). Participants were told to write a paragraph describing an episode when carefully reasoning through a situation led them in the right direction and resulted in a good outcome. Adapting this task to the online setting, we asked participants to write four sentences rather than eight-to-ten sentences as in the original task. Despite this modification, whereas at least 95% of the initially recruited participants completed the study in other conditions (i.e., answered all questions, including the survey), this figure was only 79% for R2. Among those who completed R2, the compliance rate (i.e., the prevalence of participants who wrote four or more sentences) was 88.6%. Because exclusion of non-compliant participants can jeopardize internal validity by annulling randomization (Bouwmeester et al., 2017; Tinghog et al., 2013), we include them in our analyses consistent with our preregistered intention-to-treat analysis plan.

The third reflection condition (R3) included the justification task, which elicited justifications from participants similar to Miller and Fagley (1991). Specifically, on each of the four screens where answers to the CRT-2 questions were elicited, participants were asked to justify their answers in a separate cell by providing an explanation of their reasoning in one sentence or more. For each question, the answer to the CRT-2 question and its justification were submitted simultaneously.

As the fourth reflection condition (R4), we developed a novel training task for the online context. The task was designed to improve vigilance against three commonly observed cognitive biases. Participants were asked to answer three questions. The first question was intended to illustrate a semantic illusion: "How many of each animal did Moses take on the ark?" The second question involved a test of the base rate fallacy: "In a study, 1000 people were tested. Among the participants, there were 5 engineers and 995 lawyers. Jack is a randomly chosen participant in this study. Jack is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs. What is most likely?" (Jack is a lawyer or engineer). The third question was designed to exhibit availability bias: "Which cause more human deaths?" (sharks or horses). After each question, the screen displayed the correct answer, along with an explanation of the bias (see materials at the OSF study site). Finally, participants were asked to write four sentences summarizing what they have learned in training, and they were instructed to rely on reflection during the next task (i.e., the CRT-2).

We devised a fifth reflection condition (R5) that combined decision justification (R3) with debiasing training (R4). Participants first participated in the debiasing training and then they were asked to justify their responses to the CRT-2 questions, as described above. Hence, R5 promoted learning-by-doing (Bruce & Bloch, 2012), the application of the lessons received during debiasing training on CRT-2 questions.

Two control conditions were designed to allow insightful comparisons to the five reflection conditions. The passive control condition (C1), where participants completed CRT-2 without any additional tasks, measures baseline CRT-2 scores in the participant pool. In the active control condition (C2), participants were first asked to describe an object of their choosing in four sentences before answering the CRT-2 questions. This neutral writing task in C2 controls for any direct effect that the act of writing itself in R2, R4 and R5 may have on reflection. Similarly, to achieve comparability between reflection manipulations, participants in R1 and R3 were asked to complete the same neutral writing task as in C2 prior to beginning CRT-2.

After the CRT-2, participants answered two questions on a 7-point Likert scale (1 = "not at all", 7 = "a great deal"): 1) "To what extent did you rely on your feelings or intuitions when making your decisions?", and 2) "To what extent did you rely on reason when making your decisions?" The score on the first question was reversed and the average of the scores on the two questions constituted the self-reported composite index of reflection.

Finally, participants completed a survey, including the 20-item Positive and Negative Affect Schedule (PANAS; Watson, Clark & Tellegen, 1988) and a brief demographic questionnaire. The PANAS consisted of two 10-item scales measuring positive and negative affect. Participants were
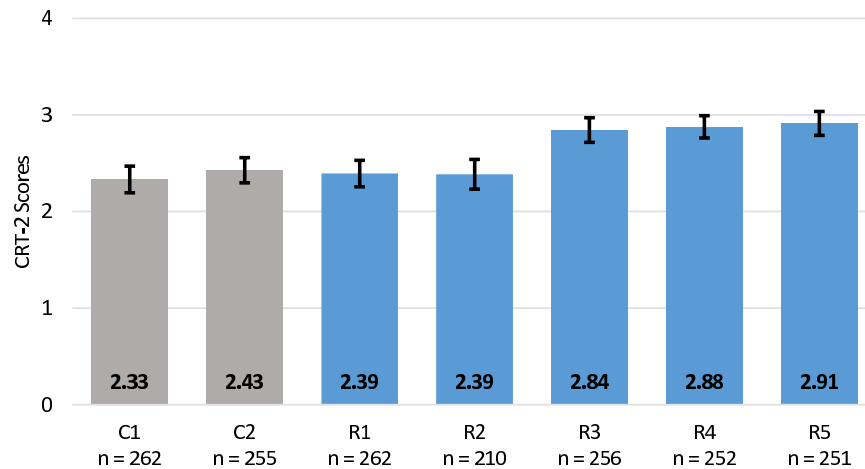
FIGURE 1: CRT-2 scores across the conditions. Sample size (n) and average number of correct answers on the Cognitive Reflection Test-2 (Thomson & Oppenheimer, 2016) in the control conditions (C1 to C2, gray bars) and the cognitive reflection manipulations (blue bars): (R1) Time delay, (R2) Memory recall, (R3) Decision justification, (R4) Debiasing training, and (R5) Debiasing training with decision justification. Error bars show 95% confidence intervals.

asked to indicate the extent to which they experienced each emotion item during the previous task (i.e., CRT-2) on a Likert scale ranging from 1 ("very slightly or not at all") to 5 ("extremely"). Both positive and negative affect scales revealed sufficient internal consistency (both Cronbach's αs = .89).

3 Results

3.1 Confirmatory tests

Overall, the debiasing training, the justification task, and their combination significantly improved performance on the CRT-2, whereas time delay and memory recall were not helpful. The CRT-2 scores across the control and experimental conditions are presented in Figure 1. A one-way ANOVA model revealed significant differences in CRT-2 scores across the conditions ($F(6, 1741) = 15.75$, $p < .001$, $\eta_p^2 = .051$). As post-hoc analysis, we conducted pairwise comparisons using two-tailed t-tests, which indicated partial support for our initial hypothesis that reflection manipulations increase performance on the CRT-2. As predicted, CRT-2 scores in the justification and debiasing training conditions (i.e., R3, R4 and R5) were significantly higher than both of the control conditions, C1 (Cohen's $d = 0.47$, 0.52 and 0.54 respectively, ps < .001) and C2 ($d = 0.40$, 0.45 and 0.47, ps < .001). In contrast, neither time delay (R1) nor memory recall (R2) showed significant difference from C1 (vs. R1: $p = .537$, $d = 0.05$; vs. R2: $p = .610$, $d = 0.05$;) or C2 (vs. R1: $p = .721$, $d = 0.03$; vs. R2: $p = .682$, $d = 0.04$). We also found partial support for our second hypothesis that debiasing training is more effective than the other reflection manipulations: CRT-2 scores in the conditions with debias-

ing training (R4 and R5) were significantly higher than time delay (R1 vs. R4: $d = 0.47$; R1 vs. R5: $d = 0.49$; ps < .001) and memory recall conditions (R2 vs. R4: $d = 0.48$; R2 vs. R5: $d = 0.50$, ps < .001) but not the justification condition (R3 vs. R4: $p = .704$, $d = 0.03$; R3 vs. R5: $p = .448$, $d = 0.07$). Failing to find confirmatory evidence for our final hypothesis, CRT-2 scores in the two conditions with debiasing training did not significantly differ (R4 vs. R5: $p = .681$, $d = 0.04$). In other words, the combination of debiasing training with justification provided no clear added benefits.

3.2 Exploratory analyses

Here, we first report the remaining (i.e., non-confirmatory) pairwise comparisons of experimental conditions, and then explore differences in response times (RTs), task noncompliance, self-reported reflection, and self-reported emotions across the conditions. No difference in CRT-2 scores were identified when comparing the two control conditions ($p = .324$) and when comparing time delay with memory recall ($p = .944$). The CRT-2 scores were higher in the decision justification condition than in the memory recall ($p < .001$). Finally, CRT-2 scores in the decision justification condition were significantly higher than the time delay condition ($p < .001$).

To help explore response times (RTs), Table 2 indicates the position of the reflection manipulations and the active controls in the study procedure as well as the mean RTs across the seven conditions. We use log-transformed RTs (base 10) to account for data skewness in all exploratory analyses that involve study duration measures. RTs in both the CRT-2 and the overall study significantly differed across conditions (CRT-2: $F(6, 1741) = 274.84$, $p < .001$, $\eta_p^2 = .486$; overall:

TABLE 2: Study configuration and response times. M denotes the position of any reflection manipulation in the study proce-
dures (i.e., before or during the elicitation of the CRT-2). AC denotes the position of any active controls (i.e., a neutral writing
task to control for the act of writing; see Method). Mean RTs (in seconds) across conditions indicate the duration of the CRT-2
task ("CRT-2"), study duration except for CRT-2 RTs ("Other"), and the total study duration ("Total").

Manipulation	Position of manipulations		Response times (sec)		
	Before CRT-2	During CRT-2	CRT-2	Other	Total
Passive control (C1)			75	183	257
Active control (C2)	(AC)		77	291	368
Time delay (R1)	(AC)	M	91	290	380
Memory recall (R2)	M		72	421	493
Decision justification (R3)	(AC)	M	250	309	559
Debiasing training (R4)	M		82	471	553
R3 + R4 (R5)	M	M	221	476	697

$F(6, 1741) = 161.26$, $p < .001$, $\eta_p^2 = .357$). As expected,
pairwise comparisons with two-tailed t-tests indicated that
eliciting justifications during CRT-2 (i.e., R3 and R5) in-
creased CRT-2 RTs compared to all other conditions (ps <
.001) and that lack of reflection manipulations or active con-
trols (i.e., C1) decreased the remaining study duration (i.e.,
excluding CRT-2 RTs) compared to all other conditions (ps
$\leq$ .001). While there was no difference between the total
study durations of R3 and R4 ($p = .889$), R1 was the fastest,
R2 was the second fastest, and R5 was the slowest reflection
condition (ps $\leq$ .001). Since careful reflection requires time,
the variation in CRT-2 scores across the conditions could in
part be driven by these RT asymmetries. Consistent with
this conjecture, a linear regression of the CRT-2 scores on
two variables that together constitute the total study duration
were both positive and statistically significant (log of total
RT on CRT-2: $\beta = 0.189$, $p < .031$, $\eta_p^2 = .003$; log of re-
maining time spent on the study: $\beta = 0.260$, $p < .034$, $\eta_p^2 =
.003$).

One reason why the time delay condition failed to sig-
nificantly activate reflection may be non-compliance with
the time-limits. In R1, 44.7% of participants failed to com-
ply with the 20-second time-limit in one or more of the four
CRT-2 questions. Similarly, 21% of participants in the mem-
ory recall condition (R2) failed to complete the study and
11.4% of participants in R2 who completed the study failed
to write at least four sentences in the memory recall task.
In principle, task noncompliance could have weakened these
reflection manipulations, since CRT-2 scores were higher
among compliant than among non-compliant participants in
both R1 (2.70 vs. 2.02, $t(260) = 5.08$, $p < .001$, $d = 0.63$) as
well as R2 (2.50 vs. 1.64, $t(208) = 3.85$, $p < .001$, $d = 0.78$).
However, these differences may also be due to participants'
thinking styles, as those who tend to be reflective (i.e., those
with higher baseline CRT-2 scores) are likely to read the

task instructions more carefully. Hence, exclusion of non-
compliant participants from the analysis can bias results by
annulling random assignment (Bouwmeester et al., 2017;
Tinghog et al., 2013), and the appropriate solution would
be to increase compliance in future studies, for example by
using forced delay in R1 and stronger monetary incentives
in R2.

Next, we explore the influence of experimental manipula-
tions on self-reported reflection (Figure 2) and affect (Figure
3). A one-way ANOVA showed that the self-reported com-
posite index of reflection significantly differed between the
conditions ($F(6, 1741) = 3.08$, $p = .005$, $\eta = .011$). Pairwise
comparisons using two-tailed t-tests revealed that partici-
pants in conditions with debiasing training (R4 and R5),
consistent with differences in CRT-2 performance, reported
relying more on reason as compared to those in the passive
control (R4 vs. C1: $p = .029$, $d = .19$; R5 vs. C1: $p = .027$,
$d = 0.20$) and the memory recall conditions (R4 vs. R2: $d
= 0.32$; R5 vs. R2: $d = 0.32$; all ps < .001). As a further
indication of the failure of the memory recall condition (R2)
in activating reflection, self-reported reflection was signif-
icantly lower in R2 as compared to the active control and
the time delay conditions (R2 vs. C2: $p = .022$, $d = 0.21$;
R2 vs. R1: $p < .001$, $d = 0.26$). No other significant differ-
ence in self-reported reflection was identified between the
experimental conditions.

One-way ANOVA models of PANAS showed significant
effect on positive affect ($F(6, 1741) = 5.25$, $p < .001$, $\eta
= .018$) but failed to show effect of conditions on negative
affect ($F(6, 1741) = 2.05$, $p = .057$, $\eta_p^2 = .007$). In particu-
lar, pairwise comparisons using two-tailed t-tests indicated
that debiasing training with decision justification (R5) sig-
nificantly increased positive affect as compared to the two
controls (R5 vs. C1: $p = .001$, $d = 0.29$; R5 vs. C2: $p < .001$,
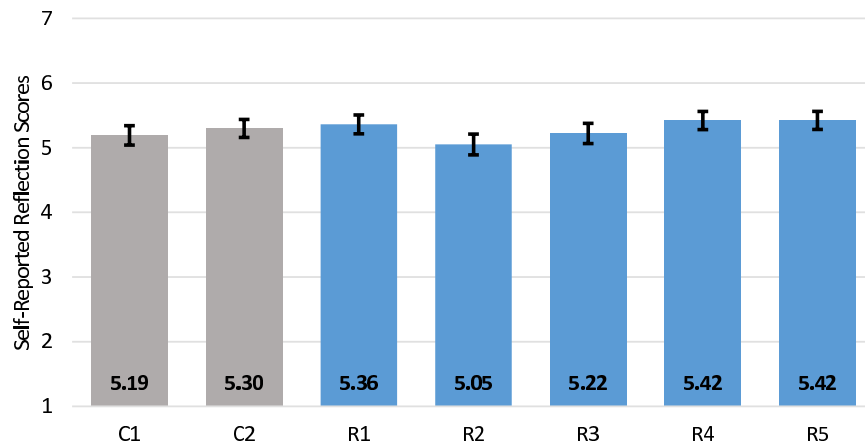$d = 0.44$) as well as the time delay (R5 vs. R1: $p = .047$, $d =$

FIGURE 2: Self-reported reflection across the conditions. Average scores on the self-reported composite index of reflection in the control conditions (C1 to C2, gray bars) and the cognitive reflection manipulations (blue bars): (R1) Time delay, (R2) Memory recall, (R3) Decision justification, (R4) Debiasing training, and (R5) Debiasing training with decision justification. Error bars show 95% confidence intervals.
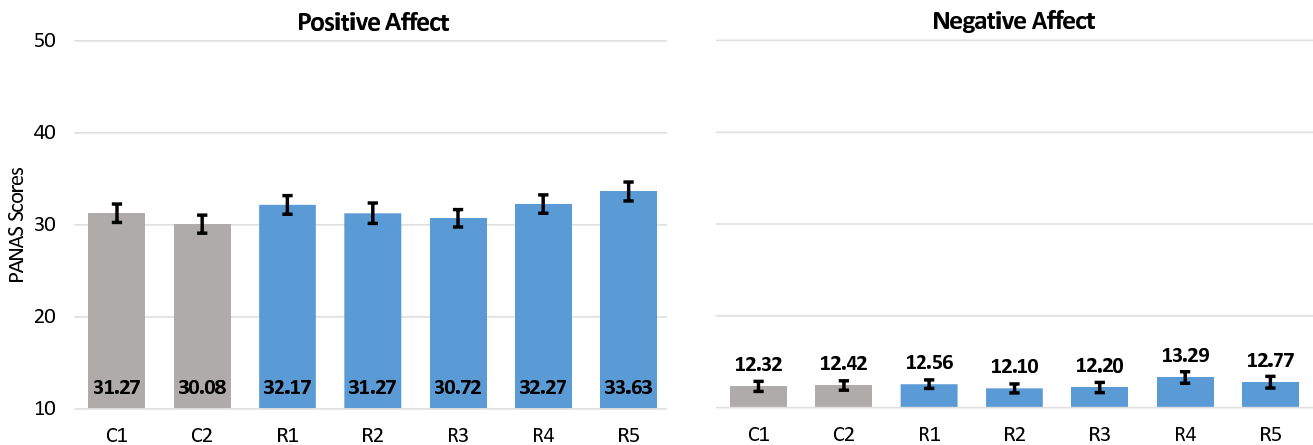

FIGURE 3: PANAS scores across the conditions. Average self-reported positive (left panel) and negative (right panel) affect scores in the control conditions (C1 to C2, gray bars) and the cognitive reflection manipulations (blue bars): (R1) Time delay, (R2) Memory recall, (R3) Decision justification, (R4) Debiasing training, and (R5) Debiasing training with decision justification. Error bars show 95% confidence intervals.

0.18), the memory recall (R5 vs. R2: $p = .002$, $d = 0.29$), and the decision justification conditions (R5 vs. R3: $p < .001$, $d = 0.36$). Time delay (R1) and debiasing training (R4) conditions also increased positive affect compared to the active control (R1 vs. C2: $p = .004$, $d = 0.26$; R4 vs. C2: $p = .002$, $d = 0.27$) and the decision justification conditions (R1 vs. R3: $p = .040$, $d = 0.18$; R4 vs. R3: $p = .027$, $d = 0.20$). All other pairwise comparisons failed to reach statistical significance.

4 Discussion

In this study, we aimed to identify experimental manipulations that can effectively activate reflective thinking. Comparing five reflection manipulations and two control condi-

tions, we found that justifying answers to the CRT-2 (R3), receiving a brief debiasing training prior to it (R4), and the combination of the two methods (R5) significantly increased reflective thinking. Against our expectations, no difference in cognitive performance was found across these three reflection manipulations. The online versions of the two manipulations commonly used in the literature — time delay (R1) and memory recall (R2) — were not found to be effective in increasing reliance on reflection, which may have been due to high noncompliance in R1 and high dropout rates in R2. On a positive note, reflection manipulations were not found to increase negative affect, and no socially desirable responding was found in these ineffective manipulations, since the self-reported reflection scores in these conditions were not higher than the controls. Overall, our study isolated two

underutilized treatments (R3 and R4) as effective reflection manipulations appropriate for the online context and indicated that the two regularly used reflection methods (R1 and R2) may not be effective with the configurations used in this study.

Are any of the successful reflection manipulations preferable to the others? Our study revealed that R3, R4 and R5 increased reliance on reflection to a similar extent — resulting in moderate effect sizes that did not significantly differ from each other. As compared to conditions with debiasing training (R4 and R5), the condition with only the decision justification task (R3) has the advantage of involving a simple prompt that is easy to administer without the need to teach explicit rules for reflection. On the other hand, compared to the conditions that use decision justification (R3 and R5), the condition with only the debiasing training (R4) achieved not only high scores but also fast responses in the CRT-2 that was subsequently elicited. Therefore, the debiasing training shows promise in inducing continued activation of reflection, but the longevity of this manipulation, as well as alternative ways to strengthen it, should be further explored. Likewise, R5 (and to a lesser extent R4) resulted in higher levels of self-reported positive affect as compared with the controls, suggesting that debiasing training and the application of its lessons during decision making can increase positive effect. Whether positive affect in turn aids reflection is an open question that needs further examination. Overall, we advise that the best reflection manipulation is the one that is most appropriate for the experimental task at hand. For example, asking justifications for decisions in tasks that measure prosocial intentions can motivate socially desirable responding. For such tasks, debiasing training can be preferable. In other research settings, decision justification can provide a fast and effective reflection manipulation.

The present study suffers from various limitations. Most importantly, our results are limited by its reliance on CRT-2 as the sole cognitive performance measure. While it is well-established that the CRT-2 scores show significant positive correlations with other cognitive reflection measures such as the CRT (Thomson & Oppenheimer, 2016; Yilmaz & Saribay, 2017c) or standard heuristics-and-biases questions (e.g., Lawson et al., 2020), it is currently unclear exactly what aspects of cognitive reflection are directly captured by the CRT-2. The CRT-2 items differ from the standard CRT items by design, relying more on careful reading than on numeracy (Thomson & Oppenheimer, 2016). In this sense, the CRT-2 items can be likened to the so-called "stumpers" (Bar-Hillel, Noah & Frederick, 2018; Bar-Hillel, Noah & Shane, 2019). On the other hand, while stumpers are difficult riddles that "do not evoke a compelling, but wrong, intuitive answer" (Bar-Hillel et al., 2018), the intuitive answers on the CRT-2 are systematically wrong and can be used to distinguish between intuitive and reflective thinking. For example, more than a third of the answers to the first CRT-2

question ("If you're running a race and you pass the person in second place, what place are you in?") in the original study by Thomson and Oppenheimer (2016) was "first" and not "second". These systematic mistakes are probably in part due to careless reading but also because correct response on this item requires the logical inference that passing the second person in a race implies the existence of another runner who is ahead of them both. Nevertheless, more research is needed to distinguish between various cognitive performance tasks in their ability to measure different aspects of reflection (e.g., Erceg, Galić & Ružojčić, 2020).

Secondly, our results are not conclusive about the potential of time delay and memory recall tasks in increasing reflection. Our setup, where the memory recall task was shortened for the online context and where the time delay condition was not forced, may have weakened the manipulations. Low task compliance in time delay and high dropout rates in memory recall could have contributed to this failure. Hence, improved methods are needed to test the superiority of the decision justification and the debiasing training tasks over time delay and memory recall. For such tests, the standard version of the memory recall that requires writing of eight sentences can be coupled with higher monetary incentives to motivate task compliance, and the alternative version of the time delay condition that forces participants to wait for a set period can be used.

Thirdly, we cannot rule out the possibility that the direct effects of our successful reflection manipulations on cognitive performance may have been limited. For example, rather than activating reflection directly, the debiasing training condition may have indirectly improved reflection performance by increasing test-taking ability through exposure to questions that are similar to the CRT-2 or by increasing understanding of the CRT-2 items through more careful reading. Likewise, the decision justification task may be open to experimenter demand effects in some contexts. One reason why we did not find evidence for socially desirable responding may be the fact that all participants were exposed to the CRT-2 prior to reporting how much they reflected. Exposure to CRT-2 may have created a sense of reliance on reflection in the control conditions. Future studies specifically designed to study the role of socially desirable responding in reflection manipulations are needed.

Overall, this study fills an important gap in the literature by highlighting two effective manipulations (and their combination) for activating reflective thinking. These methods can be easily implemented in future research on dual-process models, including experiments conducted online. Some of the commonly used reflection manipulations are recently shown to be ineffective (e.g., Deppe et al., 2015; Meyer et al., 2015), and earlier findings based on these manipulations often fail to replicate (e.g., Sanchez et al., 2017). Hence, previous results based on unreliable reflection manipulations should be tested using improved methods. Our findings indicate that,

rather than just reminding people of the benefits of reflection (as in memory recall) or giving them time to think (as in time delay), providing guidance about how to reflect specifically (as in debiasing training and decision justification) can improve cognitive performance. The methods advanced in this study — decision justification, debiasing training and their combined use — can serve this purpose well.

References

Almashat, S., Ayotte, B., Edelstein, B., & Margrett, J. (2008). Framing effect debiasing in medical decision making. *Patient Education and Counseling, 71*(1), 102–107. http://dx.doi.org/10.1016/j.pec.2007.11.004.

Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics, 21*, 99–131.

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

Bakhti, R. (2018). Religious versus reflective priming and susceptibility to the conjunction fallacy. *Applied Cognitive Psychology, 32*(2), 186–191. http://dx.doi.org/10.1002/acp.3394.

Bar-Hillel, M., Noah, T., & Frederick, S. (2018). Learning psychology from riddles: The case of stumpers. *Judgment & Decision Making, 13*(1), 112–122.

Bar-Hillel, M., Noah, T., & Shane, F. (2019). Solving stumpers, CRT and CRAT: Are the abilities related? *Judgment and Decision Making, 14*(5), 620–623.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Belardinelli, P., Bellé, N., Sicilia, M., & Steccolini, I. (2018). Framing effects under different uses of performance information: An experimental study on public managers. *Public Administration Review, 78*(6), 841–851. http://dx.doi.org/10.1111/puar.12969.

Białek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology, 28*(5), 631–639. http://dx.doi.org/10.1080/20445911.2016.1156118.

Bouwmeester, S., Verkoeijen, P., Aczel, B., Barbosa, F., Begue, L., Branas-Garza, P., . . . Wollbrant, C. E. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspect Psychol Sci, 12*(3), 527–542. http://dx.doi.org/10.1177/1745691617693624.

Bruce, B. C., & Bloch, N. (2012). Learning by doing. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1821–1824). Boston, MA: Springer US.

Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2013). When do we lie? *Journal of Economic Behavior & Organization, 93*, 258–265. http://dx.doi.org/10.1016/j.jebo.2013.03.037.

Capraro, V., Schulz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics, 79*, 93–99. http://dx.doi.org/10.1016/j.socec.2019.01.007.

Cheng, F.-F., Wu, C.-S., & Lin, H.-H. (2014). Reducing the influence of framing on internet consumers' decisions: The role of elaboration. *Computers in Human Behavior, 37*, 56–63. http://dx.doi.org/10.1016/j.chb.2014.04.015.

Christensen, J. (2018). Do justification requirements reduce motivated reasoning in politicians' evaluation of policy information? An experimental investigation. *An Experimental Investigation.(December 3, 2018)*.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *40*(2), 428–434. http://dx.doi.org/10.3758/brm.40.2.428.

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

Deck, C., Jahedi, S., & Sheremeta, R. (2017). The effects of different cognitive manipulations on decision making. *Economic Science Institute, Working Paper*.

Deppe, K. D., Gonzalez, F. J., Neiman, J. L., Jacobs, C., Pahlke, J., Smith, K. B., & Hibbing, J. R. (2015). Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology. *Judgment and Decision Making, 10*.

Erceg, N., Galić, Z., & Ružojčić, M. (2020). A reflection on cognitive reflection–testing convergent/divergent validity of two measures of cognitive reflection. *Judgment and Decision Making, 15*(5), 741–755.

Evans, J. S., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspect Psychol Sci, 8*(3), 223–241. http://dx.doi.org/10.1177/1745691612460685.

Everett, J. A. C., Ingbretsen, Z., Cushman, F., & Cikara, M. (2017). Deliberation erodes cooperative behavior — Even towards competitive out-groups, even when using a control condition, and even when eliminating selection bias. *Journal of Experimental Social Psychology, 73*, 76–81.

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

Forstmann, M., & Burgmer, P. (2015). Adults are intuitive mind-body dualists. *Journal of Experimental Psychology: General, 144*(1), 222–235.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25–42.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science, 336*(6080), 493–496. http://dx.doi.org/10.1126/science.1215647.

Gervais, W. M., van Elk, M., Xygalatas, D., McKay, R. T., Aveyard, M., Buchtel, E. E., . . . Riekki, T. (2018). Analytic atheism: A cross-culturally weak and fickle phenomenon? *Judgment and Decision Making, 13*, 268–274.

Grimm, P. (2010). Social desirability bias. In J. Sheth & N. K. Malhotra (Eds.), *Wiley international encyclopedia of marketing*. New York: John Wiley & Sons.

Horstmann, N., Ahlgrimm, A., & Glöckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. *Judgment and Decision Making, 4*(5), 335–354.

Horstmann, N., Hausmann, D., & Ryf, S. (2009). Methods for inducing intuitive and deliberate processing modes. In G. A & W. C (Eds.), *Foundations for tracing intuition: Challenges and methods* (pp. 219–237). New York, NY: Psychology Press.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399–425. http://dx.doi.org/10.1007/s10683-011-9273-9.

Igou, E. R., & Bless, H. (2007). On undesirable consequences of thinking: Framing effects as a function of substantive processing. *Journal of Behavioral Decision Making, 20*(2), 125–142. http://dx.doi.org/10.1002/bdm.543.

Isler, O., Maule, J., & Starmer, C. (2018). Is intuition really cooperative? Improved tests support the social heuristics hypothesis. *PLoS One, 13*(1), e0190560. http://dx.doi.org/10.1371/journal.pone.0190560.

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.

Klein, C. (2011). The dual track theory of moral decision-making: A critique of the neuroimaging evidence. *Neuroethics, 4*(2), 143–162. http://dx.doi.org/10.1007/s12152-010-9077-1.

Kocher, M. G., & Sutter, M. (2006). Time is money — Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization, 61*, 375–392.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nat Commun, 6*, 7455. http://dx.doi.org/10.1038/ncomms8455.

Lawson, M. A., Larrick, R. P., & Soll, J. B. (2020). Comparing fast thinking and slow thinking: The relative benefits of interventions, individual differences, and inferential rules. *Judgment and Decision Making, 15*(5), 660.

Leboeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making, 16*(2), 77–92. http://dx.doi.org/10.1002/bdm.433.

Leisti, T., Radun, J., Virtanen, T., Nyman, G., & Häkkinen, J. (2014). Concurrent explanations can enhance visual decision making. *145*, 65–74. http://dx.doi.org/10.1016/j.actpsy.2013.11.001.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131. http://dx.doi.org/10.1177/1529100612451018.

Ma, Y., Liu, Y., Rand, D. G., Heatherton, T. F., & Han, S. (2015). Opposing oxytocin effects on intergroup cooperative behavior in intuitive and feflective minds. *Neuropsychopharmacology, 40*(10), 2379–2387. http://dx.doi.org/10.1038/npp.2015.87.

Maule, A. J., Hockey, G. R. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta Psychologica, 104*(3), 283–301.

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., . . . Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *J Exp Psychol Gen, 144*(2), e16–30. http://dx.doi.org/10.1037/xge0000049.

Meyer, A., Zhou, E., & Shane, F. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making, 13*(3), 246.

Miller, P. M., & Fagley, N. S. (1991). The effects of framing, problem variations, and providing rationale on choice. *Personality and Social Psychology Bulletin, 17*(5), 517–522.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences, 14*(10), 435–440. http://dx.doi.org/10.1016/j.tics.2010.07.004.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 129–140. http://dx.doi.org/10.1177/2372732215600886.

Myrseth, K. O. R., & Wollbrant, C. E. (2017). Cognitive foundations of cooperation revisited: Commentary on Rand et al.(2012, 2014). *Journal of Behavioral and Experimental Economics, 69*, 133–138.

Neo, W. S., Yu, M., Weber, R. A., & Gonzalez, C. (2013). The effects of time delay in reciprocity games. *Journal of Economic Psychology, 34*, 20–35. http://dx.doi.org/10.1016/j.joep.2012.11.001.

Palan, S., & Schitter, C. (2018). Prolific.ac — A subject pool for online experiments. *Journal of Behavioral and*

Experimental Finance, 17, 22–27. http://dx.doi.org/10.1016/j.jbef.2017.12.004.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2013). The role of analytic thinking in moral judgements and values. *Thinking & Reasoning, 20*(2), 188–214. http://dx.doi.org/10.1080/13546783.2013.865000.

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. http://dx.doi.org/10.1016/j.cognition.2012.03.003.

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

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychol Sci, 27*(9), 1192–1206. http://dx.doi.org/10.1177/0956797616654455.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*(7416), 427–430. http://dx.doi.org/10.1038/nature11467.

Sanchez, C., Sundermeier, B., Gray, K., & Calin-Jageman, R. J. (2017). Direct replication of Gervais & Norenzayan (2012): No evidence that analytic thinking decreases religious belief. *PLoS One, 12*(2), e0172636. http://dx.doi.org/10.1371/journal.pone.0172636.

Saribay, S. A., Yilmaz, O., & Körpe, G. G. (2020). Does intuitive mindset influence belief in God? A registered replication of Shenhav, Rand and Greene (2012). *Judgment and Decision Making, 15*(2), 193–202.

Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing Training Improves Decision Making in the Field. *Psychological Science, 30*(9), 1371–1379. http://dx.doi.org/10.1177/0956797619861429.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: cognitive style influences belief in God. *J Exp Psychol Gen, 141*(3), 423–428. http://dx.doi.org/10.1037/a0025391.

Sieck, W. R., Quinn, C. N., & Schooler, J. W. (1999). Justification effects on the judgment of analogy. *27*(5), 844–855. http://dx.doi.org/10.3758/bf03198537.

Sieck, W. R., & Yates, J. F. (1997). Exposition effects on decision making: Choice and confidence in choice. *Organizational Behavior and Human Decision Processes, 70*(3), 207–219. http://dx.doi.org/10.1006/obhd.1997.2706.

Sirota, M., Theodoropoulou, A., & Juanchich, M. (2020). Disfluent fonts do not help people to solve math and non-math problems regardless of their numeracy. *Thinking*

& Reasoning, 1–18. http://dx.doi.org/10.1080/13546783.2020.1759689.

Song, H., & Schwarz, N. (2008). Fluency and the detection of misleading questions: Low processing fluency attenuates the Moses illusion. *Social Cognition, 26*(6), 791–799.

Spiliopoulos, L., & Ortmann, A. (2018). The BCD of response time analysis in experimental economics. *Experimental Economics, 21*, 383–433.

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.

Stephens, R. G., Dunn, J. C., Hayes, B. K., & Kalish, M. L. (2020). A test of two processes: The effect of training on deductive and inductive reasoning. *Cognition, 199*, 104223. http://dx.doi.org/10.1016/j.cognition.2020.104223.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition, 119*(3), 454–458. http://dx.doi.org/10.1016/j.cognition.2011.01.018.

Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition, 133*(3), 572–585. http://dx.doi.org/10.1016/j.cognition.2014.08.006.

Takemura, K. (1994). Influence of elaboration on the framing of decision. *The Journal of Psychology, 128*(1), 33–39. http://dx.doi.org/10.1080/00223980.1994.9712709.

Thompson, V. A., Evans, J., & Frankish, K. (2009). Dual process theories: A metacognitive perspective. *Ariel, 137*, 51–43.

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

Tinghog, G., Andersson, D., Bonn, C., Bottiger, H., Josephson, C., Lundgren, G., . . . Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature, 498*(7452), E1–2; discussion E2–3. http://dx.doi.org/10.1038/nature12194.

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. http://dx.doi.org/10.3758/s13421-011-0104-1.

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin, 40*(7), 923–930. http://dx.doi.org/10.1177/0146167214530436.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.

Vieider, F. M. (2011). Separating real incentives and accountability. *14*(4), 507–518. http://dx.doi.org/10.1007/s10683-011-9279-3.

Wang, C. S., Sivanathan, N., Narayanan, J., Ganegoda, D. B., Bauer, M., Bodenhausen, G. V., & Murnighan, K. (2011). Retribution and emotional regulation: The effects of time delay in angry economic interactions. *Organizational Behavior and Human Decision Processes, 116*(1), 46–54. http://dx.doi.org/10.1016/j.obhdp.2011.05.007.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol, 54*(6), 1063–1070. http://dx.doi.org/10.1037//0022-3514.54.6.1063.

Yilmaz, O., & Isler, O. (2019). Reflection increases belief in God through self-questioning among non-believers. *Judgment and Decision Making, 14*(6), 649–657.

Yilmaz, O., & Saribay, S. A. (2016). An attempt to clarify the link between cognitive style and political ideology: A non-western replication and extension. *Judgment and Decision Making, 11*, 287–300.

Yilmaz, O., & Saribay, S. A. (2017a). Activating analytic thinking enhances the value given to individualizing moral foundations. *Cognition, 165*, 88–96. http://dx.doi.org/10.1016/j.cognition.2017.05.009.

Yilmaz, O., & Saribay, S. A. (2017b). Analytic thought training promotes liberalism on contextualized (but not stable) political opinions. *Social Psychological and Personality Science, 8*, 789–795. http://dx.doi.org/10.1177/1948550616687092.

Yilmaz, O., & Saribay, S. A. (2017c). The relationship between cognitive style and political orientation depends on the measures used. *Judgment and Decision Making, 12*(2), 140–147.

Yonker, J. E., Edman, L. R. O., Cresswell, J., & Barrett, J. L. (2016). Primed analytic thought and religiosity: The importance of individual characteristics. *Psychology of Religion and Spirituality, 8*(4), 298–308. http://dx.doi.org/10.1037/rel0000095.