AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2023 The Author(s)
ISSN: 0096-3445

2023, Vol. 152, No. 7, 2052–2073
https://doi.org/10.1037/xge0001375

# Rationally Irrational: When People Do Not Correct Their Reasoning Errors Even If They Could

Miroslav Sirota, Marie Juanchich, and Dawn L. Holford
Department of Psychology, University of Essex

Why is it that sometimes people do not correct their reasoning errors? The dominating dual-process theories of reasoning detail how people (fail to) detect their reasoning errors but underspecify how people decide to correct these errors once they are detected. We have unpacked the motivational aspects of the correction process here, leveraging the research on cognitive control. Specifically, we argue that when people detect an error, they decide whether or not to correct it based on the overall expected value associated with the correction—combining perceived efficacy and the reward associated with the correction while considering the cost of effort. Using a modified two-response paradigm, participants solved cognitive reflection problems twice while we manipulated the factors defining the expected value associated with correction at the second stage. In five experiments ($N = 5{,}908$), we found that answer feedback and reward increased the probability of correction while cost decreased it, relative to the control groups. These cognitive control critical factors affected the decisions to correct reasoning errors (Experiments 2 and 3) and the corrective reasoning itself (Experiments 1, 4 and 5) across a range of problems, feedbacks, types of errors (reflective or intuitive), and cost and reward manipulations pre-tested and checked in five separate studies ($N = 951$). Thus, some people did not correct their epistemically irrational reasoning errors because they followed the instrumentally rational principle of the expected value maximization: They were rationally irrational.

*Keywords:* cognitive reflection test, dual-process theory, error correction, cognitive control, expected value of control model

*Supplemental materials:* https://doi.org/10.1037/xge0001375.supp

Prior research has documented that people are prone to committing reasoning errors, often labeled as irrational errors, because they violate normative principles of epistemic rationality such as normative theories of probabilistic or logical reasoning (e.g., Shafir & LeBoeuf, 2002). Sometimes people are unaware of their errors, but mounting evidence shows that people often sense or might even be fully aware of these errors but do not correct them (e.g., De Neys, 2012; Walco & Risen, 2017). Why do people not correct their reasoning errors once they realize that they have made a mistake? Here, we studied the motivational roots of decisions on whether to correct errors or not by testing a formalized model of mental effort allocation in executive functioning (Shenhav et al., 2013). Understanding the motivational forces behind why people do (not) correct their errors has a range of practical and theoretical implications. Regarding the practical implications, understanding the factors that lead to a higher chance of error correction would allow us to design more effective interventions aiming to enhance reasoning performance in many applied domains (e.g., Mamede et al., 2007). Regarding the theoretical implications, the current research can corroborate or falsify the model of effort allocation in reasoning and it can also extend the currently dominating models of reasoning: dual-process theories.

Dual-process theories explain reasoning as an interaction between two types of cognitive processes: intuitive (Type I) processes that are

This article was published Online First March 27, 2023.

Miroslav Sirota https://orcid.org/0000-0003-2117-9532

Dawn L. Holford is now at School of Psychological Science, University of Bristol.

For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

All data sets, pre-registration protocols, and materials used in this research are available at https://osf.io/hvqa4/

We thankfully acknowledge funding from the Experiment Psychology Society (awarded to Miroslav Sirota, Experiment 1) and the British Academy (awarded to Miroslav Sirota and Marie Juanchich, Experiments 2 and 3). The funding sources had no other involvement in this research than financial support. We declare no conflict of interest.

Miroslav Sirota served as lead for conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, visualization, and writing–original draft. Marie Juanchich served in a supporting role for funding acquisition, methodology, and resources. Dawn L. Holford served in a supporting role for methodology and resources. Miroslav Sirota, Marie Juanchich, and Dawn L. Holford contributed to writing–review and editing equally.

Correspondence concerning this article should be addressed to Miroslav Sirota, Department of Psychology, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. Email: msirota@essex.ac.uk

usually described as fast, effortless, and automatic; deliberate (Type II) processes that are usually described as slow, effortful, and reflective (De Neys & Pennycook, 2019; Evans, 2008, 2011; Kahneman, 2011; Sloman, 1996; Stanovich, 1999). Dual-process theories have successfully accounted for various effects and biases in reasoning, judgment, and decision-making (De Neys, 2006; Evans & Over, 2013; Stanovich & West, 1998; Thompson et al., 2011; Toplak et al., 2011), as well as in moral cognition (Bago & De Neys, 2019a), prosocial cooperation (Bago et al., 2021; Rand et al., 2012), magical thinking, superstitious, and paranormal beliefs (Risen, 2016; Walco & Risen, 2017), and online behavior such as spreading misinformation and fake news (Bago et al., 2020).

Studying the motivational forces behind the error correction process can extend the dual-process literature in two theoretically important ways. First, it would allow a more fine-grained understanding of the correction process that goes beyond the current focus on error detection. Prior research has enabled us to better understand error detection processes in reasoning and other higher cognition. Specifically, prior research has identified and formalized how our minds detect errors, focusing mainly on internal conflict detection (Bhatia, 2017; De Neys, 2012; De Neys et al., 2013; De Neys & Glumicic, 2008; Pennycook et al., 2015; Srol & De Neys, 2021), with theoretical discussions and empirical tests predominantly concentrated on how error detection activates deliberate processes (Ackerman & Thompson, 2017; Bago & De Neys, 2017; Banks & Hope, 2014; De Neys & Pennycook, 2019; Evans, 2008, 2019; Kahneman & Frederick, 2002; Sloman, 1996; Srol & De Neys, 2021; Travers et al., 2016). However, despite some notable exceptions (e.g., Evans, 2011), formalization and understanding of the correction processes have remained rudimentary. Indeed, at least in some versions of the dual-process theories (e.g., Kahneman & Frederick, 2002, 2005), the correction process is reduced to a step that occurs automatically if an error is detected (Patel et al., 2019; Risen, 2016). Recent experimental evidence puts into question this assumption behind the automaticity of correction processes (e.g., Bago & De Neys, 2019a, 2019b; Bago et al., 2021; Walco & Risen, 2017).

Second, our approach allows us to decompose the motivational forces behind the correction process. Motivation is known to play an important role in reasoning, based on prior research in dual-process theories. For instance, prior research found that individual differences involved in people's motivation to be rational and their need for cognition reliably predicted normative performance in a range of reasoning, judgment, and decision-making tasks (Frederick, 2005; Sirota et al., 2014; Stanovich & West, 1998, 2000; Toplak et al., 2014). These individual differences in motivation even predicted performance when cognitive abilities and executive functions were statistically controlled for (e.g., Stanovich & West, 1998). In addition, prior research found that various manipulations increasing participants' motivation to engage in more effortful thinking—for instance, direct or indirect instructions to spend more time on a task (Lawson et al., 2020; Sirota, Theodoropoulou, et al., 2021), increasing accountability by justifying answers (Isler et al., 2020; Sieck & Yates, 1997), and providing a monetary reward for accurate answers (Enke et al., 2021)—led to, at least for some types of tasks (e.g., those that did not require additional knowledge), an improved normative performance. Notwithstanding the importance of this evidence, it merely shows that motivation to engage in more effortful thinking improves reasoning. However, this research lacks any formalized decomposition of the factors contributing to the motivation to engage in the effortful activity during the error detection and error correction phase.

To formalize this decomposition process, we leveraged the current literature on motivational aspects of cognitive control using the cost and benefit associated with the effort (e.g., Kool et al., 2017; Kurzban et al., 2013; Shenhav et al., 2017). For instance, the Expected Value of Control (EVC) model explains how people become motivated to engage in a particular task, while integrating research on motivation and cognitive control into a computational and neuropsychologically implemented mechanism of cognitive control allocation (Shenhav et al., 2014, 2017, 2021). Even though the control model was initially developed and tested using less cognitively complex tasks (e.g., a Stroop task; Frömer et al., 2021; Shenhav et al., 2013) than those typically used in reasoning research (e.g., syllogistic problems), it would be parsimonious to assume that similar cognitive control allocation processes operate over higher cognitive processes (Pennycook, 2018).

In a nutshell, according to the EVC model, people decide how much and what type of effort to exert to accomplish a task by considering the chances that the allocated control will allow them to attain the desired outcome and by weighting the costs and benefits of allocating control to the task (Shenhav et al., 2021). When applied to error correction, people aim to maximize the overall expected value associated with the correction determined through the probability of correcting the error, given the effort (i.e., efficacy), expected benefits, and cost associated with the correction, which can be formalized as follows (Shenhav et al., 2013):

$$\mathrm{EVC(cs)} = \sum_{\mathrm{i}} \mathrm{P(x_i|cs)^{*}v(x_i)} - \mathrm{cost(c)}. \tag{1}$$

The decision of whether to engage in correction or not will therefore depend upon the EVC. This is a combination of the conditional probability of reaching the correct answer ($x_i$) given the exerted cognitive control ($c$) needed and given the current state ($s$) and the value, or benefits, associated with the desired state, $v(x_i)$, minus the cost associated with the exerted cognitive control ($c$). For instance, if a person has a low control efficacy, then even a relatively high benefit such as a financial reward will not result in a high control intensity (e.g., when using ice cream as a reward, I can motivate my 8-year-old daughter to correct her errors in simple multiplication problems but not complex derivation problems). A person can also decide not to correct their answer despite a relatively high control efficacy and benefit because they perceive a very high expected cost of control (e.g., even ice cream will not suffice if there are too many multiplication problems requiring substantial effort and time). Thus, this model decomposes the motivation for whether or not to correct reasoning errors. Consequently, a seemingly suboptimal engagement in correction leading to an incorrect, epistemically irrational answer could be the result of an optimal effort engagement determined by an instrumentally rational cost–benefit analysis, pitting the potential benefit of correction against its associated costs.

Prior research in reasoning offers some evidence to support the EVC model. First, negative evaluative performance feedback, which indicates to a reasoner that insufficient control intensity has been employed, leads to intensifying control and, sometimes, to improved performance (Bago et al., 2019; Ball, 2013; Ball et al., 2010). Furthermore, correlational evidence exists about the perceived control efficacy. For instance, more mathematically anxious people perform worse in numerical cognitive reflection problems (Juanchich et al., 2020; Sirota, Dewberry, et al., 2021). Finally, as we documented above, monetary and various nonmonetary rewards lead to increased effort and, under some circumstances, also to increased accuracy (Enke et al., 2021; Isler et al., 2020; Lawson et al., 2020). However, more dedicated tests of the EVC model's components are needed to better understand the motivational forces behind the error correction.

## The Present Research

In the present research, we tested whether and under which conditions people correct their reasoning errors once they have recognized them. To do so, we modified a *two-response paradigm*, originally developed to parse (more) intuitive from (more) analytical thinking (Bago & De Neys, 2017; Thompson et al., 2011). In our modified paradigm, participants answered the same reasoning problems twice. In the first stage, participants solved the reasoning problems without any time restrictions (except for Experiments 4 and 5, where time restriction was used to elicit intuitive reasoning). In the second stage, participants were *unexpectedly* asked to reconsider their initial answers to the same set of reasoning problems. We manipulated the variables of interest in the second stage of responding. Thus, our modification of this paradigm differed in three important aspects. First, our participants were not prompted to answer as quickly as possible in the first stage of responding (with the exception of Experiments 4 and 5); rather, they were provided with as much time as they needed to solve the problems. Second, our participants were not aware beforehand that they would be responding twice; they were *unexpectedly* asked to reconsider their initial answers to the same set of reasoning problems presented in the first stage. Third, our participants were not responding after each problem; rather, they saw the same problems presented in two separate blocks.

We devised several tests of the effect of feedback, reward, and cost across five experiments. Specifically, in Experiment 1, we tested whether participants could solve more problems on repeated exposure, whether answer feedback (correct vs. incorrect answer) to their initial answers improved the correction, and whether a performance-based reward improved the rate of correction. In Experiments 2 and 3, we changed the nature of the feedback, exchanging the answer feedback for explanatory feedback. The explanatory feedback allowed us to make the participants fully aware of their reasoning errors and the reasons why they were incorrect; thus, it provided the unequivocal knowledge available to correct the errors. To minimize the cognitive effort involved, we offered the participants only two options: the correct and incorrect (but default) answer. This allowed us to test the effect of additional cost (Experiments 2 and 3) and reward (Experiment 3) in the conditions where the perceived control efficacy was maximal. Finally, we tested whether reasoning cost (Experiment 4) and reward (Experiments 4 and 5) affect corrective reasoning of intuitively generated errors. Thus, in Experiments 2 and 3, we tested the effect of reward and cost on the decision-making underpinning the allocation of control (i.e., decisions to execute correction). In Experiments 1, 4, and 5, we tested the effect of reward and cost on the control over reasoning itself (i.e., the corrective reasoning).

In general, we hypothesized that our participants would assign more cognitive effort to repeated problems, resulting in higher accuracy (Hypothesis 1 in Experiment 1, hereafter, Hypothesis e1.1), and the feedback would serve as a signal to intensify mental effort resulting in higher accuracy (Hypotheses e1.2, e2.1). We also hypothesized that participants would perform a cost–benefit analysis, such that accuracy would be depreciated by cost (Hypotheses e2.2, e3.1, e4.1–e4.4) and amplified by reward (Hypotheses e1.3, e3.2, e4.1–e4.4, e5.1–e5.2). We used numerical and verbal problems of the Cognitive Reflection Test because these should be particularly sensitive to increased mental effort translating into a higher solution rate and should not require additional knowledge to solve the problems correctly (Frederick, 2005).

Finally, we also tested several predictions associated with correctness confidence for the initially incorrect trials, which were not derived from the EVC model. We reasoned that if we observed the predicted changes in accuracy, there should also be corresponding changes in meta-cognitive perception such as correctness confidence. Indeed, robust evidence indicates that reasoners' correctness confidence is sensitive to accuracy and confidence changes are observed even when reasoners do not know the normatively correct answer to the problems (e.g., De Neys, 2014; De Neys et al., 2011, 2013; De Neys & Feremans, 2013).

## Experiment 1

In a factorial design, we tested how feedback and performance-based rewards affect the solution rate of 10 open-ended cognitive reflection problems. After the first round of responding, we asked participants to solve the same problems again. In this repeated stage of responding, we manipulated whether participants received answer feedback (participants either learnt about the most common incorrect answer or not) and a performance-based reward (participants were either paid based on their performance for correct answers or not).

Guided by the EVC model, we formulated three hypotheses. First, we assumed that participants would allocate additional cognitive control to the same problems upon repeated exposure, which would translate into increased accuracy. Thus, we hypothesized that participants in the baseline condition (i.e., those who did not receive feedback or performance-based reward) would spontaneously correct some of their errors relative to the level of correction that occurred due to random error (Hypothesis e1.1). We tested the spontaneous correction conservatively: against an estimated random error of switching from a correct to an incorrect answer in the baseline condition rather than against zero. Second, we assumed that feedback would signal to the participants who made mistakes the need for allocating more cognitive control to the problems, which should translate into increased accuracy relative to the baseline condition. Therefore, we hypothesized that answer feedback (vs. no feedback) would increase the probability of correction (Hypothesis e1.2). Finally, we hypothesized that providing a reward—a performance-based monetary incentive—would increase the probability of correction (Hypothesis e1.3) because we assumed that the incentive would increase cognitive control.

We also tested two predictions associated with correctness confidence for the initially incorrect trials. Specifically, for participants who gave an initially incorrect answer, we predicted that receiving the feedback showing the incorrect intuitive answer would "shake" participants' confidence and lead to a decrease in confidence at the second stage relative to the participants without feedback (Hypothesis e1.4), whereas being rewarded (vs. not) would lead to an increase in accuracy and therefore in correctness confidence as well (Hypothesis e1.5).

## Method

### *Participants*

We recruited 1,210 participants from the online panel Prolific in exchange for a flat fee of £1.00 for a 12-min-long questionnaire. The sample size was based on an a priori stopping rule to reach the sample size of 1,200 participants; the 10 additional participants were due to the online recruiting method implemented by Prolific. The target sample size was based on a power analysis adjusted for the fact that we were interested in the correction of the initially incorrect trials. We aimed to detect a small-to-medium effect of Cohen's $f = 0.15$ using a 2 × 2 between-subjects analysis of variance (ANOVA) with 5% Alpha and 90% power (Cohen, 1988), which required 469 participants. However, the effects can be detected only in trials with incorrect responses, which we conservatively estimated to be at least 40% based on recent data (Sirota, Dewberry, et al., 2021). This means that in the most extreme scenario, 60% of participants would not make any errors. Therefore, we multiplied the resulting sample size by 2.5 (i.e., the ratio of 1/0.40). This resulted in 1,173 participants for the required sample size, which we rounded up to 1,200 to account for a possible attrition rate. Participants were eligible to take part only if they (a) had a minimal 90% approval rate in previous Prolific studies, (b) had not participated in previous studies using the Verbal or Numerical CRT run by our lab, (c) were current UK residents, and (d) were at least 18 years old.

The participants' ages ranged from 18 to 84 years ($M = 35.8$, $SD = 11.6$ years). Participants were mostly women (69.3%; 30.7% men). The sample varied in terms of education: 1.4% did not complete high school education, 38.0% completed high school, 42.2% had a college degree, 15.3% had a master's degree, and 3.1% had a PhD or other professional degree. The sample also varied in terms of occupation; the most common occupation categories were management and professionals (25.6%), followed by unemployed, students, and homemakers (24.0%), sales and office (12.7%), service (6.8%), government (4.5%), retired (3.6%), and some other less common occupations and unclassified occupations.

### *Design*

We employed a modified two-response paradigm (Thompson et al., 2011): Participants first answered 10 cognitive reflection problems and then were *unexpectedly* asked to answer the same problems again. The problems were presented in a random order to each participant in both stages. In a 2 (answer feedback: no feedback vs. feedback) × 2 (performance-based reward: no reward vs. reward) between-subjects design, we manipulated the answer feedback by presenting the most common incorrect answer and reward by providing a performance-based monetary incentive associated with correction for this second stage of answering (an additional £0.10). Participants read short instructions at the onset of the second stage of the study just before seeing the same problems again. For instance, the participants who received feedback and a performance-based reward read the following instructions:

> In the next section, we will ask you to solve the same 10 word problems again.
>
> We will give you a useful hint for solving the problem this time. In our hint, we are going to show you the most common mistake that people make with each of these problems. This means that if your previous answer matches the mistake we have provided in our hint, then you did not initially provide the correct answer to the problem.
>
> If you want to keep your previous answer, please write down the same answer in the space provided. However, if you want to change your answer, please write down the new answer in the space provided.
>
> For the next 10 word problems, you will receive a **10p bonus payment** for each correct answer that you give (in addition to your participation fee and regardless of your previous answers). Thus, if you provide correct answers to the 10-word problems you will receive your participation fee of £1 plus an extra £1 (so, £2 in total). If you do not provide correct answers to any of the word problems you will still receive your participation fee of £1 but no extra payment (so, £1 in total).

In the no-feedback conditions, participants did not see the second paragraph providing a useful hint, and the participants in the no-performance-based-reward conditions did not see the last paragraph describing the bonus payment. For each problem, the participants in the feedback conditions were provided with answer feedback (e.g., HINT: "Nunu" is NOT the correct answer) and the participants in the performance-based-reward conditions were paid a £0.10 bonus payment for each correct response. Participants were randomly allocated to the conditions by the Qualtrics built-in randomizer, which operates automatically using the Mersenne Twister algorithm (Matsumoto & Nishimura, 1998). Thus, this study was a double-blind, randomized, controlled trial.

### *Materials and Procedure*

After providing informed consent, participants answered 10 open-ended CRT problems. Five numerical cognitive reflection problems were taken from the Expanded Cognitive Reflection Test (Frederick, 2005; Toplak et al., 2014), and five problems were used from the Verbal Cognitive Reflection Test (Sirota, Dewberry, et al., 2021). The problems were presented in a random order. We measured the time taken to solve each problem. Participants also answered questions about their correctness confidence and prior familiarity with each of the questions. The confidence question ("How confident are you that the answer you have just provided is correct?") was measured on a visual analogue scale ranging from 0 to 100 (verbally anchored at 0 = *totally not sure* to 100 = *totally sure*). The familiarity question ("Have you answered this word problem prior to taking this survey?") was measured using two options (yes/no). After answering all 10 problems, we measured the perception of effort allocated to solving the problems ("How much effort did you put into solving these problems?") on a 7-point Likert scale (1 = *none at all*, 2 = *very little*, 3 = *a little*, 4 = *a moderate amount*, 5 = *a lot*, 6 = *a great deal*, 7 = *maximal possible*).

In the second stage, participants were asked to answer the same 10 problems and the associated confidence questions again. They were always reminded of their initial answer and were offered the chance to either keep it or change it. For instance, for the "Mary's father" problem, a participant in the feedback and no-performance-based-reward condition with an incorrect initial answer read:

> *Please answer the same word problem again. We will give you a hint this time.*
>
> Mary's father has 5 daughters but no sons – Nana, Nene, Nini, Nono. What is the fifth daughter's name probably?
>
> Your previous answer was: **Nunu**
>
> **HINT: "Nunu" is NOT the correct answer**
>
> *If you want to keep your previous answer, please write the same answer below.*
>
> *If you want to change your previous answer, please write the new answer below.*

The participants in the no-feedback conditions did not see the answer-specific feedback (i.e., the "hint") and were only reminded of their initial answer. Finally, the participants in the reward conditions saw the same materials but were told to be rewarded for each correct response (see exact wording of all materials used in the Online Supplemental Material B).

After completing the problems in the second stage, the participants answered a subjective effort question concerning this second stage with identical wording to the first stage. They were also asked whether they used any external help to answer the problems during the second stage ("In this last round of answering, have you tried to answer these word problems by looking at some external sources [e.g., the internet, books, asking friends]?") using a 6-point Likert scale (0 = *none of them*, 1 = *very few of them*, 2 = *few of them*, 3 = *some of them*, 4 = *almost all of them*, 5 = *all of them*). To encourage honest reporting of their efforts, participants were reassured that they would receive their payment—a flat participation fee and, if applicable, a performance-based fee—regardless of their answer to that question. Finally, the participants completed a medical probability judgment task unrelated to this research, answered some socio-demographic questions, and were debriefed.

### Transparency and Openness

We conducted all studies presented in this manuscript in accordance with the ethical standards of the American Psychological Association and obtained ethical approval from the Ethics Committee of the Department of Psychology of the University of Essex. We have reported how we determined our sample size, all measures, manipulations, and exclusions in all of the studies. Experiments 1, 4, and 5 were not pre-registered. Experiments 2 and 3 were pre-registered. The materials, data sets, and pre-registrations are publicly available on the Open Science Framework at https://osf.io/hvqa4/ (Sirota et al., 2022).

## Results

### Manipulation Checks

We first evaluated the effect of feedback and reward manipulation on the subjective mental effort reported in Stage 2 (adjusted for baseline effort reported in Stage 1) and the average time spent on the problems in Stage 2 (adjusted for the baseline time taken in Stage 1) (see Table 1). Participants allocated more cognitive control in the conditions with feedback and reward. The subjective mental effort increased when the Stage 2 problems were accompanied by the feedback relative to the mental effort reported in Stage 1, whereas it decreased when Stage 2 did not feature any feedback, $F(1, 1{,}204) = 12.02$, $p < .001$, Cohen's $f = 0.10$. The subjective mental effort increased when the performance was rewarded, relative to the baseline effort, and decreased when it was not, $F(1, 1{,}204) = 27.51$, $p < .001$, Cohen's $f = 0.15$. Their interaction between feedback and reward was not significant, $F(1, 1{,}204) = 0.82$, $p = .364$, Cohen's $f = 0.03$. The average time spent on each problem overall decreased in the second stage, but it decreased less when the problems were accompanied by the feedback, $F(1, 1{,}206) = 36.50$, $p < .001$, Cohen's $f = 0.17$, and when participants were rewarded for correct solutions, $F(1, 1{,}206) = 9.54$, $p = .002$, Cohen's $f = 0.09$. Their interaction between feedback and reward was not significant, $F(1, 1{,}206) = 0.78$, $p = .377$, Cohen's $f = 0.03$. The observed differences in time spent on each problem according to the feedback could be explained by longer text (seven words more, ∼10% longer) in the feedback conditions relative to other conditions. However, this only partially explains the substantial 8–10 s difference between the conditions.

### Effect of Feedback and Reward on the Probability of a Correction

On average, participants correctly answered slightly more than a third of the problems in the first stage, with negligible variation across the conditions (see Table 1). Participants improved their performance in the second stage, but we also observed differences according to the manipulated conditions (see Table 1). To test our model-derived hypotheses, we calculated the probability of correction on the initially incorrect trials—Only 16 participants did not make any initial errors. By correction, we mean when a participant changed their response from an initially incorrect answer to a correct one (rather than making changes to the initially incorrect answer). We used all of the initially incorrect trials rather than only the initially intuitively incorrect trials expressed in the feedback for two reasons. First, the non-intuitively incorrect answers represented a non-negligible source of errors (2,589 incorrect non-intuitive responses represent 34% of all incorrect responses); their removal would substantially decrease the statistical power. Second, for most of the items the non-intuitive errors were quite similar to the intuitive errors (e.g., "Nuno," "Nonu," etc. instead of "Nunu") or blatantly incorrect (e.g., "don't know," "no idea"); their removal would thus obscure the rate of spontaneous correction and the effect of the reward. In addition, 153 participants reported that they used some external help at least for one of the questions (one of the 16 participants reported using external help in the second stage). This affected the probability of correction: Participants who reported using external help ($M = 0.30$, $SD = 0.33$, $n = 153$) managed to correct their answers more often than participants who did not use external help ($M = 0.20$, $SD = 0.26$, $n = 1{,}042$), $t(2{,}315.4) = 6.81$, $p < .001$, Cohen's $d = 0.38$. Therefore, we removed the data from these participants. These two exclusion steps still left us with sufficient power to test our hypotheses ($n = 1{,}042$).

The participants in the condition without feedback and performance-based rewards tended to spontaneously correct their initial errors more than estimated by random error, $t(237) = 7.96$, $p < .001$, Cohen's $d = 0.52$ (Figure 1). As predicted, the rate of correction varied according to the conditions (Figure 1). The participants who received answer feedback were more likely to correct their errors than those without feedback, $F(1, 1{,}038) = 50.07$,

**Table 1**

*Effect of Feedback and Reward on Subjective Effort, Time Spent on Problems, Correctness, and Correctness Confidence*

| | No performance-based reward | | Performance-based reward | | |
|---|---|---|---|---|---|
| | No feedback | Feedback | No feedback | Feedback | Total |
| Variables | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) |
| | | | All trials | | |
| Subjective effort change | −0.13 (0.85) | −0.02 (0.72) | +0.06 (0.72) | +0.25 (0.79) | +0.04 (0.78) |
| Mean time change (s) | −20.3 (28.0) | −12.9 (26.6) | −17.1 (24.0) | −7.3 (19.2) | −14.4 (25.1) |
| Correct responses (%) | | | | | |
| First stage | 39.7 (27.2) | 35.9 (25.9) | 36.7 (26.8) | 35.8 (26.6) | 37.0 (26.6) |
| Second stage | 45.7 (28.4) | 46.8 (29.3) | 44.7 (29.8) | 51.1 (30.0) | 47.0 (29.4) |
| Confidence (0–100) | | | | | |
| First stage | 72.9 (17.0) | 72.4 (17.9) | 71.4 (18.4) | 73.7 (16.5) | 72.6 (17.5) |
| Second stage | 78.4 (17.1) | 72.0 (21.1) | 75.8 (19.6) | 72.5 (21.1) | 74.6 (19.9) |
| | | | Incorrect valid trials | | |
| *P*(correction) | .13 (.21) | .23 (.27) | .16 (.24) | .27 (.28) | .20 (.26) |
| Confidence change | +7.0 (13.1) | −1.2 (19.2) | +6.7 (13.0) | −2.25 (21.7) | +2.6 (17.6) |

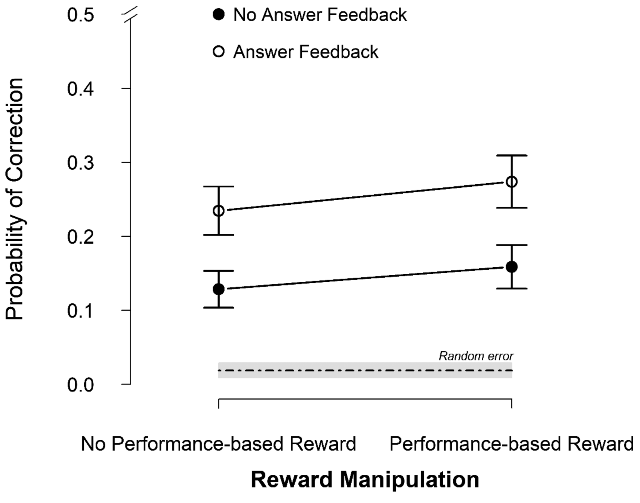
*Note.* All trials: $N = 1{,}210$; incorrect valid trials: $N = 1{,}042$.

$p < .001$, Cohen's $f = 0.22$. Those who received the reward for correct responses tended to correct their errors more than those who did not, $F(1, 1{,}038) = 4.95$, $p = .026$, Cohen's $f = 0.07$. Feedback and reward did not interact, $F(1, 1{,}038) = 0.08$, $p = .774$, Cohen's $f = 0.01$.

To further test the robustness of our conclusions regarding the effects of feedback and reward on the probability of correction, we conducted a trial-level analysis using a logistic mixed-effects model. We transformed the data into a long format and created a correction variable (0 = *not successful*/1 = *successful correction*) for each problem that had an initially incorrect answer. We ran a model with a random intercept within participants and CRT items and with the two manipulations as fixed effects (a full random-structure model did not converge). The feedback and reward increased the probability of correction, $OR = 1.57$, 95% CI [1.40, 1.76], $z = 7.67$, $p < .001$; $OR = 1.17$, [1.04, 1.31], $z = 2.63$, $p = .009$, respectively, without evidence of interaction, $OR = 1.01$, [0.90, 1.13], $z = 0.14$, $p = .886$. Hence, using a trial-level analysis corroborated our hypotheses that feedback and performance-based reward motivated participants to correct their initially incorrect answers. Thus, the results supported our model-derived Hypotheses e1.1, e1.2, and e1.3.

**Figure 1**

*Effects of Feedback and Reward on the Probability of Correcting an Originally Incorrect Response*



*Note.* $N = 1{,}042$; circles represent means; error bars represent 95% confidence intervals. A random error of correction: The dot-dashed line indicates the mean of the estimated random error of correction, which was calculated as the mean level of correction from correct to incorrect responses in the no-feedback and no-performance-based-reward condition ($n = 244$); the shaded area represents 95% confidence intervals of the estimated random error.

### *Effect of the Feedback and Reward on Correctness Confidence*

On average across the problems, participants expressed relatively high confidence with the answers they provided in the first stage, but without any remarkable differences in their confidence across the conditions (see Table 1). Participants' confidence slightly increased in the second stage, but the direction and size of the change depended on the feedback (see Table 1). To test our hypotheses concerning relevant confidence changes, we subtracted the initial confidence from the confidence in the second stage for problems where participants gave an initial incorrect answer and did not seek external help ($N = 1{,}042$). As predicted, participants who received feedback showed a slight decrease in confidence, whereas those who did not receive feedback had an increase in confidence; the difference between these two feedback groups was statistically significant, $F(1, 1{,}038) = 65.35$, $p < .001$, Cohen's $f = 0.25$. Contrary to our prediction, however, the rewards did not cause any notable differences in confidence, $F(1, 1{,}038) = 0.39$, $p = .534$, Cohen's $f = 0.02$. The interaction term was also not significant, $F(1, 1{,}038) = 0.11$, $p = .735$, Cohen's $f = 0.01$. Thus, only one hypothesis (Hypothesis e1.4) but not the other (Hypothesis e1.5) was supported.

## Discussion

Overall, we found support for the predictions of the EVC model (Shenhav et al., 2016): The answer feedback and performance-based-rewards motivated participants to allocate more cognitive control when given the problems again and, in turn, increased the probability of correcting their mistakes. Participants in the no-reward and no-feedback condition also tended to correct some of their errors above and beyond the estimated random correction error. Hence, at least some participants could have provided a correct answer in the first round of answering if they had been motivated to allocate more cognitive effort to solving the problems.

However, we did not manipulate control efficacy and the cognitive cost of correction, which are critical components of the EVC model. Even though we excluded those who reported that they searched for the answer online at least for one problem, we cannot completely exclude the possibility that some participants did not disclose that they searched for the answers online. We addressed these limitations in the subsequent experiments by adjusting the two-response paradigm so that it eliminated any motivation to search for the answers online while testing the effect of efficacy, reward, and cognitive cost.

## Experiment 2

In Experiment 2, we aimed to further test the EVC model while overcoming the limitations of Experiment 1. To do so, we redesigned the cognitive reflection problems used in Experiment 1 in a way that would allow participants to indisputably detect the errors in their prior reasoning and offered the opportunity to correct these errors while imposing on participants different levels of costs associated with the correction.

First, participants were asked to solve cognitive reflection problems with two-answer options instead of open-ended problems and those in the feedback condition were provided with explanatory feedback that featured the correct answer and a detailed explanation of why the correct answer was correct and the intuitive answer was incorrect. These two changes effectively brought participants' control efficacy—the first term of Equation 1, the conditional probability of reaching the correct answer ($x_i$) given the exerted control ($c$) needed and given the current state ($s$)—close to 1 (this was indeed the case as shown in Pre-test 1, see the online supplemental material A). Second, participants were either asked (or not) to complete an additional task if they attempted to correct their initial answer. This allowed us to test the effect of the cognitive cost linked to correction but dissociated from the control efficacy linked to the primary cognitive reflection problems. In other words, we externalized the cognitive cost associated with the correction using a novel and pre-tested cognitive cost task (see Pre-test 2, online supplemental material A).

Strictly speaking, such externalizing of the cognitive cost means that the designed task represents two control allocation events—the control allocated to the decision whether to correct a wrong answer or not and the control over categorization itself. Each of these processes could be decomposed using the EVC model. Thus, we do not claim that the allocated control studied here is only control over reasoning, as was the case in Experiment 1. However, we construe this task as *a joint control event*. In deciding whether or not to correct their answer, all participants had to allocate some control to the process of correction by evaluating the explanation and justification of the correct answer. In some conditions, participants also had to allocate some control to the categorization task. To decide whether the correction was worth it, participants therefore had to weigh up the benefits of correction (e.g., the value attached to being correct and the monetary value for being correct in some conditions). They also had to weigh up the costs associated with such correction (e.g., intrinsic cost to evaluate the justification; the cost associated with the cognitive effort and time invested in completing the additional [categorization] task). In such a case, this joint control event goes beyond simple action selection (i.e., whether to proceed with the categorization task or not) since participants undergo a cost–benefit analysis of correction. Indeed, in a simple conceptualization check, we found that if participants ($N = 401$) were asked only to choose between two otherwise indifferent options, with only one involving completing a categorization task, 21% selected such a high-cost option (i.e., the decision to allocate control over categorization). However, if the decision also involved providing a correct answer to a reasoning problem, as was used in Experiment 2, 68% selected such an option, which is significantly more than if they were simply choosing between actions, $\chi^2(1) = 88.83$, $p < .001$, Cramer's $V = 0.47$ (see the Online Supplemental Material A). Thus, participants weighed up the costs and benefits of the categorization task as well as the costs and benefits of the correction decision.

The EVC model predicts that when people perceive a very high (close to 1) probability of providing the correct response while encountering minimal cost they will correct their reasoning errors. Thus, we expected that participants would correct their reasoning errors more when provided with explanatory feedback compared with those without feedback (Hypothesis e2.1). Furthermore, we expected that participants would correct their reasoning errors less often when the costs associated with the correction were increased compared with those who only faced baseline costs (e.g., costs associated with response switching; Hypothesis e2.2). Finally, we hypothesized the interaction between cost and feedback. Specifically, we expected a stronger effect of increased costs (vs. baseline costs) in the feedback conditions than in the no-feedback conditions (Hypothesis e2.3). Such an interaction effect hinges on the assumption that people are not sufficiently capable of correcting their errors without feedback and thus their lack of ability obscures the depreciating effect of cognitive costs.

We also tested three predictions associated with changes in correctness confidence between the two stages. We assumed that any changes in problem-solving accuracy would manifest in changes in correctness confidence. First, we predicted that people would have a higher increase in their correctness confidence when provided with explanatory feedback compared to no feedback (Hypothesis e2.4). Second, we predicted that people would have a lower increase in their correctness confidence when facing an increased cost compared to a baseline cost (Hypothesis e2.5). Finally, we predicted that people would have a higher increase in their correctness confidence with feedback (vs. no feedback) while facing only a baseline cost compared with facing an increased cost (Hypothesis e2.6).

### Method

#### *Participants*

We recruited 1,753 participants from the online panel Prolific in exchange for a flat fee of £1.30 for a 15-min questionnaire. The sample size was based on an a priori stopping rule to reach the sample size of 1,750 participants; the three additional participants were due to the online recruiting method implemented by Prolific. The target sample size was based on a power analysis, which assumed to detect a small effect of Cohen's $f = 0.1$ in 2 × 2 ANOVA, which yielded $n = 787$ (Cohen, 1988). The resulting number was adjusted for the fact that such an effect would be detected only in the trials with initially incorrect responses. In our pre-test we found, on average, 55% correct responses to the cognitive reflection problems with the two-answer options we used here (see Pre-test 1, the online supplemental material A). In the most extreme scenario, this would mean that 55% of participants would not make any errors, so we multiplied the sample size by 1/0.45, resulting in 1,750 (1,749 rounded up). Participants were eligible to take part only if they had a minimal 80% approval rate in previous studies, did not participate in the previous studies using the Verbal CRT run by our lab, were UK nationals and current UK residents and were at least 18 years old. Some participants were excluded automatically if they were timed out by the system or if they failed three manipulation checks (see the online supplemental material B). We did not exclude anybody for responding too quickly.

The participants' ages ranged from 18 to 82 years ($M = 34.4$, $SD = 12.5$ years); 62.5% were women, 37.1% were men, and 0.5% were of other gender identity. The sample varied in terms of the highest achieved education: 1.3% did not complete their high school education, 37.4% completed high school education, 43.9% completed a college degree, 14.3% completed a master's degree, and 3.1% completed a PhD or other professional degree. The sample also varied in terms of occupation; the most common occupation category was management and professionals (28.5%) followed by students (15.7%), unemployed (9.7%), sales and office (9.0%), government (5.4%), service (5.0%), retired (3.9%), and some other less common occupations such as construction and production workers and unclassified occupations.

### Design

We employed a two-response paradigm similar to Experiment 1: The participants first answered the cognitive reflection problems and then were unexpectedly asked to answer the same problems again. In a 2 (feedback: no feedback vs. explanatory feedback) × 2 (cost: baseline cost vs. increased cost) between-subjects design, we manipulated the presence of explanatory feedback and increased cost associated with correction in the second stage of answering. In the feedback and increased-cost condition, participants read the following instructions:

> In the next section, we will ask you **to solve the same eight word problems again.**
>
> This time, **we will give you feedback on your previous answer**: we will tell you whether your answer was correct or not and explain why this is the case. You will have the opportunity to provide a new answer to the same problem, so if you initially made an incorrect answer then you should change your initial answer.
>
> **If you decide to change your original answer you will then have to complete a word categorisation task each time you decide to change your original answer.**
>
> Before we start, we are going to ask you to complete an example of the word categorisation task on the next page.

The participants in the baseline-cost conditions did not see the sentence in the third paragraph (about the additional task), whereas the participants in the no-feedback conditions did not see the first sentence of the second paragraph (about getting feedback). For each problem, the participants in the explanatory-feedback conditions were provided with complete feedback on their answer (e.g., "Your original answer was 'Nunu.' This is incorrect. The correct answer is 'Mary.'") and the explanation for the correct answer (see the example in Materials and Procedure below). The participants in the cost conditions had to solve an additional cost task if they wished to change their original answer (see the example below). As in Experiment 1, this study was a double-blind randomized controlled trial with automatic randomization to the conditions.

Before running the study, we pre-tested whether the explanatory feedback was effective. We asked 100 participants (77.0% women, 22.0% men, and 1.0% other gender; with ages ranging from 19 to 67 years, $M = 34.5$, $SD = 10.9$ years) on Prolific to complete the cognitive reflection problems presented in a modified two-response paradigm. The feedback was effective: After receiving the explanatory feedback, the initial solution rate ($M = 0.56$, $SD = 0.23$) increased substantially ($M = 0.97$, $SD = 0.08$), $t(99) = -18.71$, $p < .001$, $d = -1.87$; it was also indistinguishable from the maximal performance assuming a random correction error (i.e., 0.97), $t(99) = -0.47$, $p = .636$, $d = -0.05$. The random error (0.03) was determined based on the observation that 3 out of 100 participants mentioned distraction/an answer error as the reason for not correcting their reasoning error (see Pre-test 1, the online supplemental material A).

### Materials and Procedure

After providing informed consent, participants answered eight of the cognitive reflection problems used in Experiment 1. The problems featured a multiple-choice answer format with two-answer options: the correct answer and the "intuitively appealing" incorrect answer (Sirota & Juanchich, 2018). We excluded two problems from the set of problems used in Experiment 1 because we found that the solution rate in the "match" problem was too high in the first stage already and in the "sell-a-pig" problem the solution rate did not reach a maximal level even after explaining the feedback (see Pre-test 1 in the online supplemental material A). The problems and answer options were presented in a random order. For each item, participants also evaluated their confidence in their answers as in Experiment 1. Participants were also exposed to three instructional manipulation check questions: Two were transformed from the unused cognitive reflection problems asking participants to select one of two answer options and were intermixed with other problems in the first stage. The third one occurred after answering all the problems in the first stage and instructed the participants to select a number 11 out of a set of five numbers.

Participants were then unexpectedly asked to answer the same eight problems and the associated confidence questions again. They were always reminded of their initial answer and were offered the chance to either keep or change it. However, the instructions differed according to the conditions to which the participants were allocated. For instance, for the "Mary's father" problem, the participants with an incorrect initial answer in the feedback and increased-cost condition read:

> Please answer the same word problem again.
>
> *The problem was:* Mary's father has 5 daughters but no sons – Nana, Nene, Nini, Nono. What is the fifth daughter's name probably?

> A/ Nunu
>
> B/ Mary
>
> *Feedback:* Your original answer was “Nunu.” This is incorrect. The correct answer is “Mary.”
>
> *Explanation:* Here we explain why the fifth daughter’s name is Mary.
>
> The father has four daughters whose names all begin with N: Nana, Nene, Nini and Nono, but also we learned at the beginning of the first sentence that he is Mary’s father. Since Mary is a female name, his fifth daughter must be Mary.
>
> The answer “Nunu” is a common incorrect response that people give when they overlook the fact that the father is also Mary’s father.
>
> If you want **to keep** your original answer, please select “Nunu” from the options below.
>
> If you want **to change** your original answer, please select “Mary” from the options below.
>
> **Note:** If you do decide to change your original answer you will have to complete a version of the word categorisation task you went through earlier in order to make the change.

The participants in the no-feedback conditions did not see the answer feedback and the explanation of the answer—only the reminder of their initial answer. The participants in the baseline-cost conditions did not see the reminder of the cost associated with changing the answer (i.e., the note at the very bottom of the instructions). Again, the answer options (e.g., “Nunu,” “Mary”) were randomized; the presentation order of the problems was randomized as well. The participants who answered the initial set of problems correctly received the same information, but the feedback said that the answer they provided was correct (see the online supplemental material B for complete wording of the materials).

Just before answering the problems in the second stage, all participants completed a word-categorization task to experience the time and mental effort needed to complete the task. This was done because the participants in the increased-cost conditions who decided to change their original answer in the second round of answering were asked to complete one of the four variants of the word-categorization task (inspired by Heled et al., 2012). Participants thus experienced the cognitive effort of categorization. In the task, participants had to sort 10 words (e.g., “under,” “pool,” “mother”) into three categories: (a) five letters and at least two vowels; (b) four letters and at least two vowels, and (c) all other words. To proceed to the next question, participants had to categorize all the words correctly. We measured the time they took to complete the task. Every time they completed it, they were reminded that they had to complete the task because they wanted to change their answer. We developed and pre-tested this cost task in two phases. In Phase 1, participants ($N = 251$, 64.5% women, 35.1% men, and 0.4% other gender; with ages ranging from 18 to 74 years, $M = 29.9$, $SD = 10.5$ years) assessed three sorting tasks with increased difficulty—using one, two, and three sorting dimensions to consider as the categorization rules for easy, medium, and hard sorting tasks, respectively—and syllogistic problems measuring belief bias as a benchmark task since these are often substituted for cognitive reflection problems (Baron et al., 2015; Markovits & Nantel, 1989). Using two sorting dimensions of categorization rules (i.e., the medium level of difficulty with sorting according to the number of letters and vowels) was the best match with the belief bias syllogistic problem in terms of the subjective and behavioral measures of the cognitive costs (e.g., Cooper-Martin, 1994; Dunn et al., 2016; Pollock et al., 2002). In Phase 2, participants ($N = 99$, 63.6% women, 34.3% men, and 2.0% other gender identity; with ages ranging from 18 to 69 years, $M = 34.1$, $SD = 12.3$ years) completed the medium sorting task using a list with 10 and 20 words as well as a cognitive reflection problem. The 10-word list categorization task was the closest match with a cognitive reflection problem in terms of cognitive cost (for more details, see Pre-test 2, online supplemental material A).

After completing the second stage of responding, participants who did not correct their initially incorrect answer at least once saw a set of four follow-up questions probing their reason for non-correction. The reasons suggested in the questions were either associated with the control effectivity (e.g., “I did not know how to correct my answer”) or the perceived cost (e.g., “I thought it would take too much mental effort to correct the answer”). The questions were presented randomly and measured on a 7-point Likert scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *somewhat disagree*, 4 = *neither agree nor disagree*, 5 = *somewhat agree*, 6 = *agree*, 7 = *strongly agree*). Finally, the participants completed a task unrelated to this research, answered a socio-demographic question, and were debriefed.

## Results

### *Effect of the Explanatory Feedback and Correction Cost on the Probability of Correction*

On average, in the first stage of answering, participants correctly solved more than half of the problems with no noticeable differences across the conditions (Table 2). In the second stage, participants correctly answered, on average, around 21% more problems than in the first stage. However, they did so with obvious differences across the conditions—unsurprisingly, the most striking improvement was observed when they received the explanatory feedback (Table 2). To test our model-derived hypotheses, we calculated the probability of correction on the initially incorrect trials after excluding 143 participants who did not make any initial errors, which still left us with sufficient power to test our hypotheses ($N = 1{,}610$).

We found that, on average, the participants tended to correct their initial errors (Table 2). They corrected their errors in the baseline-cost and no-feedback condition above and beyond the estimated random error of correction (see Figure 2). More importantly, the probability of correction visibly depended on the manipulation of feedback and cost: When the participants received explanatory feedback, they were very likely to correct their errors but their probability of correcting their errors dropped when they faced an increased cost of correction (see Figure 2). The inferential tests confirmed these observations. As predicted, the effect of the explanatory feedback (vs. no feedback) statistically significantly increased the probability of correction, with a large effect size, $F(1, 1{,}606) = 2{,}891.69$, $p < .001$, Cohen’s $f = 1.34$. The effect of the increased cost (vs. baseline cost) statistically significantly decreased the probability of correction, with a medium effect size, $F(1, 1{,}606) = 25.99$, $p < .001$, Cohen’s $f = 0.18$. Finally, as predicted, the interaction was—despite its small effect size—statistically significant, $F(1, 1{,}606) = 5.88$, $p = .015$, Cohen’s $f = 0.06$. In Figure 2, we can see that the interaction was driven by a more pronounced effect of the

**Table 2**
*Effect of Feedback and Cost on Correctness and Correctness Confidence*

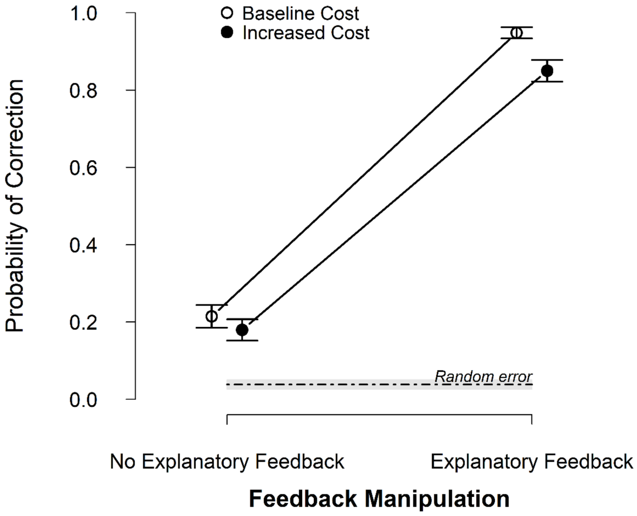
| | Baseline cost | | Increased cost | | |
|---|---|---|---|---|---|
| | No feedback | Feedback | No feedback | Feedback | Total |
| Variables | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) |
| | | | All trials | | |
| Correct responses (%) | | | | | |
| First stage | 58.8 (25.4) | 56.8 (26.7) | 57.8 (25.7) | 56.9 (25.9) | 57.6 (25.9) |
| Second stage | 64.9 (25.6) | 96.8 (9.7) | 62.3 (25.9) | 92.9 (17.4) | 78.9 (25.9) |
| Confidence (0–100) | | | | | |
| First stage | 82.8 (12.7) | 81.1 (13.5) | 81.4 (12.8) | 81.5 (13.1) | 81.7 (13.0) |
| Second stage | 87.2 (12.2) | 96.7 (8.4) | 87.8 (12.0) | 93.2 (14.7) | 91.2 (12.6) |
| | | | Incorrect valid trials | | |
| *P*(correction) | .21 (.30) | .95 (.15) | .18 (.28) | .85 (.29) | .55 (.44) |
| Confidence change | +6.5 (16.5) | +22.4 (23.4) | +10.4 (18.1) | +14.0 (30.0) | +13.3 (23.4) |

*Note.* All trials: $N = 1,753$ ($n$ per condition in the order listed in the table: $n = 437$, $n = 437$, $n = 442$, $n = 437$); incorrect valid trials: $N = 1,610$ ($n$ per condition in the order listed in the table: $n = 398$, $n = 399$, $n = 406$, $n = 407$).

cost on the probability of correction in the feedback conditions $F(1, 1,606) = 28.33$, $p < .001$ (using the Holm adjustment), Cohen's $d = 0.43$, relative to the effect of the increased cost in the no-feedback conditions, $F(1, 1,606) = 3.57$, $p = .059$ (using the Holm adjustment), Cohen's $d = 0.12$.

We conducted a pre-registered trial-level analysis using a logistic mixed-effect model, which confirmed the robustness of the main and interaction effects. We ran a model with a random intercept for participants and items and with the two factors and their interaction (using deviation coding) as fixed effects since the maximal random-structure model was not able to be estimated. The feedback increased the probability of correction, $OR = 16.56$, 95% CI [13.41, 20.46], $z = 26.03$, $p < .001$, and the cost decreased the probability of correction, $OR = 0.64$, [0.55, 0.73], $z = -6.31$, $p < .001$. The interaction between feedback and cost was significant, $OR = 0.79$, [0.69, 0.91], $z = -3.33$, $p < .001$. To unpack the interaction, we conducted simple effects analyses of the cost. The simple effect of cost in the no-feedback condition was significant (the estimated marginal means in the baseline and increased cost conditions: $M = 0.11$, $SE = 0.01$; $M = 0.07$, $SE = 0.01$, respectively), $OR = 0.81$, [0.67, 0.97], $z = -2.30$, $p = .022$. The effect of cost was also significant in the feedback condition but the effect was more pronounced (the estimated marginal means in the baseline and increased-cost conditions: $M = 0.98$, $SE < 0.01$; $M = 0.93$, $SE = 0.01$, respectively), $OR = 0.50$, [0.41, 0.62], $z = -6.37$, $p < .001$. Therefore, using a trial-level analysis corroborated our model-testing hypotheses assuming the effect of feedback and cost and their interaction. In contrast with the main analysis, the simple effect of cost on probability of correction in the no-feedback conditions was also statistically significant. Overall, these results supported our model-derived Hypotheses e2.1, e2.2, and e2.3.

**Figure 2**
*Effects of Feedback and Cost on the Probability of Correcting an Originally Incorrect Response*



*Note.* $N = 1,610$; circles represent means; error bars represent 95% confidence intervals. A random error of correction: The dot-dashed line indicates the mean of the estimated random error of correction, which was calculated as the mean level of correction from correct to incorrect responses in the no-feedback and baseline-cost condition ($n = 437$); the shaded area represents 95% confidence intervals of the estimated random error.

## *Effect of the Feedback and Cost on the Correctness Confidence*

Across all items, on average, participants expressed relatively high confidence with the answers they provided in the first stage without any remarkable differences across the conditions (see Table 2). Participants' confidence substantially increased in the second stage of answering but the direction and size of the change depended on the feedback (see Table 2).

To test our hypotheses concerning relevant confidence changes, we subtracted participants' initial confidence from their confidence in the second stage for the problems tested in prior hypotheses ($N = 1,610$). As predicted, participants who received the full explanatory feedback had a greater increase in confidence than those who did not receive the feedback, $F(1, 1,606) = 74.94$, $p < .001$, Cohen's $f = 0.22$. We also found that those with the increased cost had a lesser increase in confidence than those who did not face the increased cost, $F(1, 1,606) = 3.89$, $p = .049$, Cohen's $f = 0.05$.

The interaction between feedback and cost was statistically significant, $F(1, 1{,}606) = 29.57$, $p < .001$, Cohen's $f = 0.14$. As shown in Table 2, unpacking the interaction revealed that cost had an opposite effect on participants' confidence, as a function of people's self-efficacy. Therefore, when participants received explanatory feedback, imposing the additional cost led to lower increased confidence $F(1, 1{,}606) = 6.00$, $p = .014$, Cohen's $d = -0.23$; without the feedback, however, imposing the additional cost led to higher increased confidence, $F(1, 1{,}606) = 27.49$, $p < .001$, Cohen's $d = 0.31$ (both using the Holm adjustment). This pattern reversal can be explained through different levels of uncertainty as to the correct answer: Only relatively confident participants decided to change their answer in the no-feedback conditions, which resulted in a higher confidence increase, whereas the participants who did not correct themselves in the explanatory feedback conditions depreciated their correctness confidence because they knew they were wrong. Thus, all three hypotheses (Hypothesis e2.4–e2.6) were supported.

### *Secondary Analyses: Effect of the Feedback and Cost on the Reasons for Not Correcting*

In this pre-registered secondary analysis, we investigated how participants mentally represented the reasons for not correcting their answers. To do so, we investigated the reasons why participants did not correct their initially wrong answer using a set of follow-up questions; only those participants who did not correct at least one of their initially incorrect responses rated these scales. The structure and relative endorsement of the reasons differed across the conditions (Table 3). The dominant reason in the no-feedback conditions was the belief that the incorrect answer was not incorrect, whereas in the feedback conditions the reasons were mixed. In the feedback condition with baseline cost, the main reason for not correcting one's answer was the belief that the incorrect answer was not incorrect, whereas in the feedback condition with the increased cost, the main reasons were that it would cost too much time and mental effort.

We expected that participants would rate a lack of knowledge and belief in the correctness of their mistakes higher in the no-feedback conditions relative to the feedback conditions. We also expected that they would rate time and mental effort as the reason for not correcting in the cost conditions more than in the no-cost conditions. First, the participants in the no-feedback conditions agreed more that they lacked knowledge, $t(341.17) = 3.10$, $p = .002$, Cohen's $d = 0.21$, and that they did not make an initial mistake, $t(341.17) = 12.45$, $p < .001$, Cohen's $d = 1.60$, as the reasons for not correcting their initial errors than those in the feedback conditions. In the feedback conditions, however, we also noticed that the belief that there was no initial mistake appeared to be stronger in the baseline relative to the increased-cost condition (Table 3). Second, the participants in the baseline-cost condition agreed less with "too much time," $t(820.67) = -8.44$, $p < .001$, Cohen's $d = -1.15$, and "too much mental effort," $t(896.27) = -5.64$, $p < .001$, Cohen's $d = -0.88$, as the reasons for not correcting their initial errors than those in the increased-cost conditions. Yet, again, we noticed that the effect of cost on agreement with too much time and too much mental effort appeared to be more pronounced in the feedback condition relative to the no-feedback condition (Table 3).

**Table 3**
*Effect of Feedback and Cost on the Reasons for Not Correcting Their Errors*

| | Baseline cost | | Increased cost | | |
|---|---|---|---|---|---|
| | No feedback | Feedback | No feedback | Feedback | Total |
| Reasons | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) |
| Didn't know how | 2.2 (1.9) | 1.7 (1.1) | 2.2 (1.8) | 1.9 (1.3) | 2.1 (1.7) |
| It was not wrong | 6.2 (1.1) | 5.1 (2.0) | 6.1 (1.2) | 3.3 (2.3) | 5.7 (1.7) |
| Too much time | 1.6 (1.1) | 1.8 (1.3) | 1.9 (1.3) | 4.2 (2.2) | 2.0 (1.6) |
| Too much mental effort | 1.7 (1.2) | 2.2 (1.8) | 1.9 (1.4) | 3.5 (2.1) | 2.0 (1.6) |

*Note.* All participants with at least one no-correction trial: $n = 916$ ($n$ per condition in the order listed in the table: $n = 369$, $n = 57$, $n = 379$, $n = 111$); all the follow-up measures used a scale ranging from 1 = *strongly disagree* to 7 = *strongly agree*.

To sum up, participants' reasons for not correcting their mistakes reflected our manipulations. When participants did not have explanatory feedback, they did not correct their mistakes mainly because they believed their initial wrong answers were correct; when provided with feedback, they believed this much less. Participants also took into account the costs of correcting their answer (i.e., time and mental effort) when the increased cost was experimentally induced and it became the dominant reason for not correcting the answer in the feedback condition.

## Discussion

In Experiment 2, we confirmed the critical predictions of the EVC model. According to the model, a higher perceived probability of reaching the correct answer given the mental effort (i.e., efficacy) should manifest in the increased probability of error correction. To optimize the expected value of correction, however, any additional cognitive control should depreciate such probability of correction. Indeed, when participants' efficacy approached maximal values, they corrected their reasoning errors almost perfectly. However, even when participants were fully aware of their reasoning errors and knew how to correct them, they were less likely to correct their errors if the additional cognitive cost was experimentally imposed. Interestingly, the effect of increased cost was much more pronounced in the full explanatory conditions. This was also manifested in the confidence ratings and the reasons expressed for not correcting their answers. One limitation of Experiment 2, however, was that we did not test the effect of reward. The effect of reward would be important to test again given the limitations surrounding the effect of reward identified in Experiment 1. Besides, from a model-testing perspective, it would be important to test the effect of reward and cost jointly. We therefore addressed these limitations in Experiment 3.

## Experiment 3

In Experiment 3, we tested the joint effect of two terms in the EVC model: the reward and the cost associated with the correction. This contrasts with the previous experiments, where these two terms were tested separately. To do so, we always provided full explanatory feedback to our participants. We created three conditions: a baseline-cost condition, an increased-cost condition, and an increased-cost condition accompanied by a performance-based reward. This was preferred to a full factorial design of the cost and reward because a ceiling effect prevented us from providing a fair test of the additional effect of the reward in the baseline-cost condition. (Indeed, using the data of Experiment 2, where the mean solution rate in the explanatory feedback condition was $M = 0.95$, 95% CI [0.92, 0.97], the mean of this new condition would have to be roughly 1 to demonstrate a statistically significant increase of the reward in this condition.) Testing the effect of the reward with full explanatory feedback eliminates the motivation for searching for correct answers online (as some participants did in Experiment 1) because we provided participants with the correct answer and its explanation.

We pre-registered two model-testing hypotheses. First, we predicted that participants would correct their reasoning errors less often when the costs associated with the correction were increased compared with those with baseline costs and no reward (Hypothesis e3.1). Second, we predicted that participants would correct their reasoning errors more often when the costs associated with the correction were increased but the reward for correction was increased as well, compared with those with increased costs and no reward (Hypothesis e3.2).

We also had expectations for associated changes in correctness confidence. We predicted that people would have a lower increase in their correctness confidence in the increased-cost condition relative to the baseline condition (Hypothesis e3.3). We also predicted that people would have a higher increase in their correctness confidence in the increased-cost and reward condition relative to the increased-cost condition (Hypothesis e3.4).

## Method

### Participants

We recruited 1,452 participants from the online panel Prolific in exchange for a flat fee of £1.30 for a 15-min questionnaire. The sample size was based on an a priori stopping rule to reach the sample size of 1,449 participants; the three additional participants were due to the online recruiting method implemented by Prolific. The target sample size was based on a power analysis, which yielded $n = 652$ to detect the effect of Cohen's $d = 0.22$ (50% size of the effect of the cost on correction found in the feedback conditions of Experiment 2) in the planned contrasts with 5% Alpha and 80% power (Cohen, 1988) using an independent samples *t*-test. The resulting number was adjusted for the fact that such an effect would be detected only in the trials with initially incorrect responses (i.e., multiplied by 1/0.45). Thus, the final estimated sample size was 1,449. Participants were eligible to take part only if they had a minimal 80% approval rate in previous studies, did not participate in previous studies using the Verbal CRT run by our lab, were UK nationals and current UK residents, and were at least 18 years old. Some participants were automatically excluded if they were timed out by the system or failed three instructional manipulation checks (see Online Supplemental Material B). We did not exclude any participants for too quick responses when they completed the questionnaire. We removed one participant who completed the questionnaire twice.

The participants' ages ranged from 18 to 81 years ($M = 34.8$, $SD = 12.5$ years); 66.4% were women, 33.1% were men, and 0.5% were of other gender identity. The sample varied in terms of the highest achieved education: 1.2% did not complete their high school education, 38.1% achieved high school education, 45.6% achieved a college degree, 12.6% achieved a master's degree, and 2.5% achieved a PhD or other professional degree. The sample also varied in terms of occupation; the most common occupations were management and professionals (30.5%), students (14.9%), sales and office (9.8%), unemployed (8.6%), service (5.7%), government (5.0%), retired (4.6%), and some other less common occupations such as farmers or construction workers and unclassified occupations.

### Design

Using the two-response paradigm of Experiment 2, the participants received the feedback and effort manipulations in the second stage of answering. In a simple between-subjects design, participants were allocated to one of three conditions: (a) the baseline condition—with baseline cost and no performance-based reward, in which the correction involved no additional cost and no additional reward, (b) the increased-cost condition—with an increased cost and no performance-based reward, in which the correction involved an increased cost induced by an additional categorization task but no additional reward, and (c) the increased-cost and performance-based-reward condition—with an increased cost and also an additional reward, in which the correction involved an increased cost induced by an additional categorization task and an additional reward of £0.20 for each correction of an initially incorrect answer. The reward compensated participants for roughly 2.5 min of their time when using the minimal flat fee on Prolific (£5 per hour), which was substantially more than they would spend solving a single word-categorization task (*Mdn* = 1.3 min; see Pre-test 2, Online Supplemental Material A). The exact wording of the manipulation was slightly changed compared to Experiment 2. The participants in the increased-cost-and-reward condition saw the Stage 2 instructions as described below:

> In the next section, we will ask you **to solve the same eight word problems again.**
>
> This time, **we will give you feedback on your previous answer**: we will tell you whether your answer was correct or not and *explain* why this is the case. You will have the opportunity to provide a new answer to the same problem, so if you initially made an incorrect answer then you can change your initial answer.
>
> **However, if you decide to change your original answer you will then have to complete a word categorisation task each time you decide to change your original answer. This will cost you extra time and effort.**
>
> **You will get extra £0.20 for correcting your originally incorrect response (i.e., as a bonus payment).**
>
> Before we start, we are going to ask you to complete a word categorisation task on the next page.

In the baseline condition, the participants read the instructions above but without the bolded sentences in paragraphs three and four (explaining the additional cost and reward). In the increased-cost condition, participants read the sentences in paragraph three (explaining the additional cost) but not four (explaining the reward). For each problem, the participants in all the conditions were provided with explanatory feedback (e.g., "Your original answer was 'Nunu.' This is incorrect. The correct answer is 'Mary.'") and the explanation for the correct answer (see the example below). For each problem, the participants in the cost conditions were reminded that they would have to complete an additional categorization task if they changed their initial answer. Finally, the participants in the reward condition were reminded that for each problem they would receive a bonus payment of £0.20 for each correction and received the bonus payment after completing the study. Again, this study was a double-blind randomized controlled trial because Qualtrics automatically randomly allocated participants to the conditions.

### Materials and Procedure

We used the same materials and procedure as employed in Experiment 2. In brief, after providing informed consent, participants answered eight cognitive reflection problems with two answer options and the associated confidence questions (as well as three instructional manipulation check questions). Participants were then unexpectedly asked to answer the same eight problems and the associated confidence questions again. They were assigned to one of the three conditions and all answered the same word-categorization task as in Experiment 2 to experience the time and mental effort needed to complete the task. When answering the problems, they were always reminded of their initial answer and were offered the chance to either keep or change it. The instructions differed according to the condition to which the participants were allocated. For instance, for the "Mary's father" problem, the participants with an incorrect initial answer in the increased-cost-and-reward condition read:

> Please answer the same word problem again.
>
> *The problem was:* Mary's father has 5 daughters but no sons – Nana, Nene, Nini, Nono. What is the fifth daughter's name probably?
>
> A/ Nunu
>
> B/ Mary
>
> *Feedback:* Your original answer was "Nunu." This is incorrect. The correct answer is "Mary."
>
> *Explanation:* Here we explain why the fifth daughter's name is Mary.
>
> The father has four daughters whose names all begin with N: Nana, Nene, Nini and Nono, but also we learned at the beginning of the first sentence that he is Mary's father. Since Mary is a female name, his fifth daughter must be Mary.
>
> The answer "Nunu" is a common incorrect response that people give when they overlook the fact that the father is also Mary's father.
>
> If you want **to keep** your original answer, please select "Nunu" from the options below.
>
> If you want **to change** your original answer, please select "Mary" from the options below.
>
> **Note: If you do decide to change your original answer you will have to complete a version of the word categorisation task you went through earlier in order to make the change.**
>
> **Note: If you correct your initially incorrect response, you will receive extra £0.20 as a bonus payment.**

Participants in the baseline condition did not see the reminder of the costs associated with changing the answer or the reward reminder, whereas those in the increased-cost-but-no-reward condition saw only the cost reminder. Again, in both stages of answering, the answer options (e.g., "Nunu," "Mary") were randomized and the presentation order of the problems was randomized as well. The participants who answered the initial set of problems correctly received the same information, but the feedback said that the answer they provided was correct (see Online Supplemental Material A). The participants in the increased-cost condition who decided to change their original answer in the second round of answering were asked to complete one of the four variants of the word-categorization task (the same variants as in Experiment 2; see Online Supplemental Material B). To proceed to the next question, they had to categorize all the words correctly. The participants in the reward condition were rewarded with an extra £0.20 for each corrected answer. In contrast with Experiment 2, the participants did not answer any follow-up questions. The participants then entered an unrelated second part of the experiment, which was designed to be much shorter for the participants who were assigned to the conditions with correction costs than for the participants without them. This was done to secure fair compensation for all participants. Finally, the participants answered socio-demographic questions and were debriefed.

## Results

### Effect of the Cost and Reward on the Probability of Correction

On average, in the first stage of answering, participants correctly solved more than half of the problems with no noticeable differences across the conditions (Table 4). In the second stage, participants mostly corrected their errors upon receiving the explanatory feedback, but we also observed more pronounced differences between the conditions (Table 4). To test our model-derived hypotheses, we calculated the probability of correction on the initially incorrect trials—119 participants did not make any initial errors, which still left us with sufficient power to test our hypotheses ($N = 1{,}333$).

We found that, on average, the participants' probability of correcting their initial errors was very high, but it also varied according to the

**Table 4**
*Effect of Cost and Reward on Correctness and Correctness Confidence*

| | Baseline cost | Increased cost | | Total |
|---|---|---|---|---|
| Variables | No reward<br>*M* (*SD*) | No reward<br>*M* (*SD*) | Reward<br>*M* (*SD*) | *M* (*SD*) |
| | | All trials | | |
| Correct responses (%) | | | | |
| First stage | 58.4 (25.5) | 57.8 (26.6) | 55.1 (26.2) | 57.1 (26.1) |
| Second stage | 97.1 (08.9) | 90.9 (19.1) | 96.4 (10.6) | 94.8 (13.9) |
| Confidence (0–100) | | | | |
| First stage | 81.5 (14.7) | 81.4 (13.1) | 81.9 (13.1) | 81.6 (13.7) |
| Second stage | 96.9 (08.3) | 92.3 (18.4) | 96.2 (11.0) | 95.1 (13.4) |
| | | Incorrect valid trials | | |
| *P*(correction) | .95 (.17) | .84 (.30) | .94 (.20) | .91 (.23) |
| Confidence change | +22.2 (23.7) | +12.4 (35.2) | +20.0 (24.7) | +18.2 (28.6) |

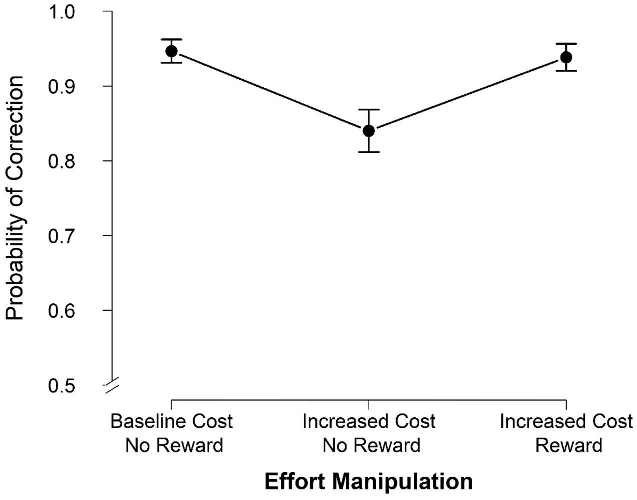
*Note.* All trials: $N = 1{,}452$ ($n$ per condition in the order listed in the table: $n = 484$, $n = 484$, $n = 484$); incorrect valid trials: $N = 1{,}333$ ($n$ per condition in the order listed in the table: $n = 441$, $n = 440$, $n = 452$).

conditions (see Table 4). The participants without the increased cost associated with the correction were very likely to correct their errors, but their probability to correct their errors dropped when they faced an increased cost. The reward (i.e., a performance-based payment) associated with correction compensated for this increased cost (see Figure 3). The observed differences were confirmed when we ran inferential statistical tests. The omnibus ANOVA was statistically significant, $F(2, 1{,}330) = 29.61$, $p < .001$, Cohen's $f = 0.21$, and it was followed by two planned contrasts to test our focal hypotheses. First, as predicted, the effect of the increased cost statistically significantly decreased the probability of correction relative to the baseline cost, with a medium effect size, $F(1, 879) = 42.01$, $p < .001$, Cohen's $f = 0.22$. Second, as predicted, the effect of reward statistically significantly increased the probability of correction relative to no reward in the increased-cost condition, with a medium effect size, $F(1, 890) = 33.40$, $p < .001$, Cohen's $f = 0.19$ (Figure 3).

Similar to the previous experiments, we conducted a pre-registered trial-level analysis using a logistic mixed-effect model, which confirmed the robustness of these model-testing conclusions. We used a mixed-effect logistic regression to analyze the effect of the manipulation following a similar logic behind the two critical planned contrasts. To test the first contrast, we ran a model with random intercepts for participants and items and with the manipulation as a fixed effect (a maximal random-structure model returned a singular fit warning). The increased cost significantly decreased the probability of correction, $OR = 0.51$, 95% CI [0.34, 0.78], $z = -3.15$, $p = .002$. To test the second contrast, we ran a model with a random intercept within participants and items and with the manipulation as a random effect. Reward in the increased-cost conditions increased the probability of correction, $OR = 2.95$, [1.24, 6.98], $z = 2.45$, $p = .014$. Thus, both model-testing hypotheses postulating the effect of cost (Hypothesis e3.1) and reward (Hypothesis e3.2) on the probability of correction were corroborated.

**Figure 3**

*Effects of Cost and Reward on the Probability of Correcting an Originally Incorrect Response*



*Note.* $N = 1{,}333$; circles represent means; error bars represent 95% confidence intervals. The probability of correction ranges from 0 to 1. Reward/No Reward refers to the presence/absence of a performance-based reward.

### *Effect of the Correction Cost and Reward on Correctness Confidence*

On average, participants expressed relatively high confidence in the answers they provided in the first round (around 82%) but again without any remarkable differences in their confidence across the conditions (Table 4). Participants' confidence increased in the second stage (around 95%), and we observed some differences between the conditions (see Table 4).

To test our pre-registered hypotheses concerning relevant confidence changes, we subtracted the initial confidence from the confidence after the feedback for the items tested in prior hypotheses (i.e., initially incorrect items) in the sample of participants who made at least one initial error ($N = 1{,}333$). Participants showed increased confidence after receiving feedback and to what extent it increased depended on the condition to which they were allocated. The omnibus ANOVA was statistically significant, $F(2, 1{,}330) = 14.66$, $p < .001$, Cohen's $f = 0.15$—two planned contrasts tested our focal hypotheses. First, as predicted, the correctness confidence increase was statistically significantly lower in the increased-cost (no reward) condition relative to the baseline cost, $F(1, 879) = 23.72$, $p < .001$, Cohen's $f = 0.16$. Second, as predicted, the correctness confidence increase was statistically significantly higher for those who received a reward relative to no reward despite the imposed additional cost of correction, $F(1, 890) = 14.08$, $p < .001$, Cohen's $f = 0.13$. The size of these effects was small to medium. Thus, both hypotheses—the negative effect of cost on confidence increase (Hypothesis e3.3) and the positive effect of reward on confidence increase (Hypothesis e3.4)—were confirmed.

## Discussion

In Experiment 3, we found further support for the critical predictions of the EVC model. While keeping the probability of reaching the correct answer close to 1 and constant across the conditions, we demonstrated the deteriorating effect of cost and the facilitating effect of reward on the probability of a correction. A joint test of the effect of cost and reward allowed us to provide direct evidence for the compensatory effect of reward on correction imposed by increased cost. Furthermore, the observed changes in correction behavior were mirrored by changes in participants' confidence in the quality of their answers: The increase in confidence was lower when correction came at a cost, except if the cost was mitigated by a reward.

# Experiment 4

Experiments 2 and 3 focused on whether or not participants decided to correct their reasoning errors, but did not focus directly on the control exerted over the reasoning itself, as was the case in Experiment 1. Therefore, we designed Experiment 4, where participants tried to solve the cognitive reflection problems that they did not initially get right, as in Experiment 1, while overcoming the limitations of Experiment 1. First, we used a new, pretested set of cognitive reflection problems that were not searchable on the internet. Thus, any effect of reward could not be ascribed to external help. Second, we used the original (unmodified) version of the dual-response paradigm, eliciting the intuitive reasoning first and then the correction phase involving mental effort (Bago & De Neys, 2017; Thompson et al., 2011). The intuition was elicited by instructions and strict limitations on the reading and reasoning time. We expected this to lead to participants making more errors and accepting that they did make a mistake more readily than in the paradigm used in prior experiments. Finally, in the second stage of responding, we asked participants to complete the cognitive reflection problems along with a simple counting task within a 40-s time limit. The counting task required participants to count instances of the letter A in a 10 × 10 matrix and served as a good benchmark for the involvement of Type II processes since it required focused attention and working memory (Healy & Nairne, 1985; Logie & Baddeley, 1987). This allowed us to manipulate the reward (by rewarding reasoning accuracy) and cost (by rewarding counting accuracy at the expense of reasoning accuracy) of reasoning. We created three conditions: a baseline condition, a cost condition manipulated by a performance-based reward given for performance in the competing task, and a reward condition manipulated by a performance-based reward given for performance in the cognitive reflection task.

Aligned with the EVC model, we expected that the cost and reward involved in the task would determine the effort allocated to the reasoning task, thus affecting people's probability of correcting their reasoning errors (Hypothesis e4.1) and counting accuracy (Hypothesis e4.3) relative to the baseline. We also hypothesized that the cost and reward would have opposite effects on the probability of correction (Hypothesis e4.2) and counting accuracy (Hypothesis e4.4).

## Method

### *Participants*

We recruited 1,101 participants from the online panel Prolific in exchange for a flat fee of £1 for a 10-min questionnaire. The sample size was based on an a priori stopping rule to reach the sample size of 1,100 participants; the additional participant was due to the online recruiting method implemented by Prolific. The target sample size was based on a power analysis. To detect a small effect of Cohen's $f = 0.12$ in a one-way ANOVA assuming $\alpha = 0.05$ and $\beta = 0.05$ and three groups, we needed $N = 1{,}077$ ($N = 906$ for each contrast with two groups). The resulting number was adjusted upwards to adjust for attrition rate—we assumed that only a few participants would not make any errors since the first answers were severely time-restricted. Participants were eligible to take part only if they had a minimal 80% approval rate in previous studies, had not participated in the previous studies using the CRT run by our lab, were UK nationals, and were at least 18 years old. Four participants were excluded due to their IP addresses being identical. The analytical sample was 1,097.

The participants' ages ranged from 18 to 87 years ($M = 41.7$, $SD = 14.5$ years); 49.8% were women, 49.6% were men, and 0.6% were of other gender identities. The sample varied in terms of the highest achieved education: 1.8% did not complete their high school education, 35.6% completed high school education, 46.3% completed a college degree, 12.9% completed a master's degree, and 3.4% completed a PhD or other professional degree. The sample also varied in terms of occupation; the most common occupation category was management and professionals (31.2%) followed by sales and office (11.1%), retired (9.1%), unemployed (8.6%), students (7.8%), government (6.2%), service (5.9%), and some other less common occupations such as construction and production workers and unclassified occupations.

### *Design*

We used a slightly modified two-response paradigm (Thompson et al., 2011): The participants first answered six cognitive reflection problems intuitively, and then, they were unexpectedly asked to answer the incorrectly answered problems again. In the first stage of responding, intuitive reasoning was elicited by imposing a strict time limit (around 10 s; see precise time for each problem in Pretest 3, Online Supplemental Material A). Participants completed all the problems in the first stage before moving to the second stage, where they were given the same problems again. In the second stage, participants were told that each of the cognitive reflection problems would be accompanied by a counting task presented on the same page and that they would have 40 s to complete both of them. Within this time limit, they could allocate as much time as they wanted to the word problem or the counting task. For this second stage of responding, we allocated participants to one of the three groups in a between-subjects design: baseline condition, cost condition, and reward condition. In the baseline condition, participants were encouraged to answer both the word problem and the counting task correctly. In the cost condition, the participants were told that they would be paid £0.40 for answering each counting task correctly. The cost manipulation induced an opportunity cost for participants if they exerted effort on the cognitive reflection problem at the expense of the counting task. The manipulation also induced cognitive cost associated with the decision whether to prioritize the counting or reasoning task and possible switching between the tasks. In the reward condition, the participants were told that they would be paid £0.40 for answering each word problem correctly (thus incentivizing correction of the cognitive reflection problem). For instance, in the reward condition, the participants read the instructions below:

> **Try your best to answer the word problem and the counting task correctly.**
>
> **You will get an extra £0.40 for answering the word problem correctly (i.e., as a bonus payment).**
>
> So if you are presented with all six word problems and answer each of them correctly, you will receive £2.40 as a bonus payment.

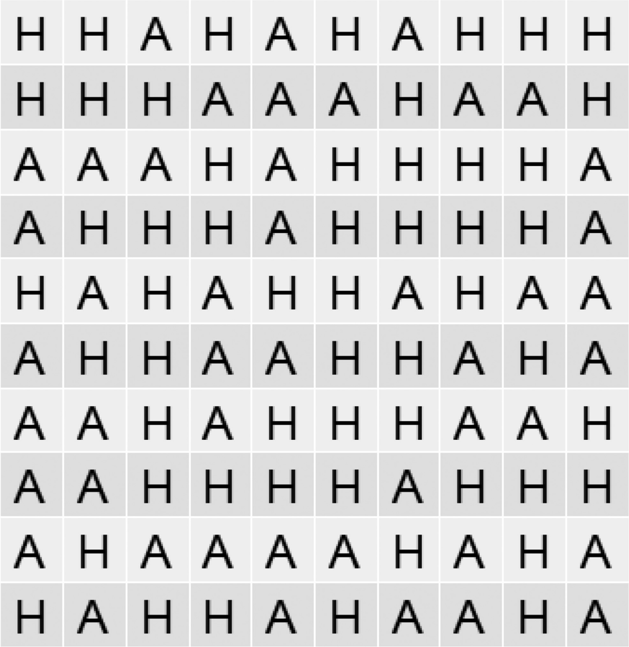
The experiment was a double-blind randomized controlled trial with automatic randomization of the conditions.

### *Materials and Procedure*

After providing informed consent, participants provided their intuitive answers to six cognitive reflection problems. The content of the problems was modified so that it would be very difficult and time-consuming to find the solution via an internet search. For instance, the widget problem was modified to this self-driving car problem: "A car factory in Dongguan has 100 new industrial robot arms that can manufacture 100 self-driving cars in 100 min. How many minutes must 200 industrial robots take to manufacture 200 self-driving cars?" All problems featured a multiple-choice answer format with four-answer options: the correct answer, the "intuitively appealing" incorrect answer and two other incorrect answers (Sirota & Juanchich, 2018). The problems and answer options were presented in a random order to each participant. To elicit intuitive answers, we put the participants under time pressure by giving them just enough time to read the problems rather than time to think about them. We determined the reading times in a pretest in which we asked 100 participants (50.0% women, 50.0% men; with ages ranging from 18 to 89 years, $M = 41.0$, $SD = 13.8$ years) on Prolific to read each problem in an online questionnaire (see Pretest 3 in the online supplemental material A for more details). Time pressure is a well-proven method for eliciting intuitive responses, and it is as effective as time pressure combined with a working memory load generated by a concurrent task (Bago & De Neys, 2017, 2019b).

Participants were then unexpectedly informed that they would again be required to answer the problems that they had not answered correctly (one at a time, presented in a random order). In addition, they were asked to complete a counting task. In the counting task, they were shown a table with 10 columns and 10 rows (i.e., 100 cells), with each cell containing a letter "A" or "H." For an example, see Figure 4. The participants had to count how many letter "As" there were. The six matrices were designed to have a letter "A" appearing in each matrix approximately but not exactly 50 times (i.e., 54, 46, 56, 48, 53, 47). Participants had to answer each cognitive reflection problem (that was previously incorrect) as well as this counting task within a 40-s limit. Both tasks were presented on the same page, and participants could see a countdown timer. This time limit was fixed, so participants could not progress more quickly; after 40 s, they were automatically moved to the next page. Based on the pretest, the time limit of 40 s should have been sufficient for participants to solve the cognitive reflection problem correctly if they decided to allocate the time to it or to complete the concurrent counting task, but not both (see Pretest 3 in the online supplemental material A). Participants solved a set of practice problems and were then allocated to one of the three conditions: baseline condition, cost condition, and reward condition (see instructions above). Before seeing a new set of tasks, they were reminded of the instructions, and the word problem and counting tasks were clearly identified in each pair of the tasks to avoid possible confusion.

**Figure 4**
*An Example of the Counting Task*



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| H | H | A | H | A | H | A | H | H | H |
| H | H | H | A | A | A | H | A | A | H |
| A | A | A | H | A | H | H | H | H | A |
| A | H | H | H | A | H | H | H | H | A |
| H | A | H | A | H | H | A | H | A | A |
| A | H | H | A | A | H | H | A | H | A |
| A | A | H | A | H | H | H | A | A | H |
| A | A | H | H | H | H | A | H | H | H |
| A | H | A | A | A | A | H | A | H | A |
| H | A | H | H | A | H | A | A | H | A |

Finally, the participants completed a short questionnaire unrelated to this research, answered a socio-demographic question, and were debriefed.

## Results

### *Effect of the Cost and Reward on the Probability of Correction*

On average, in the first intuitive stage of answering, participants correctly solved around one-third of the problems with no noticeable differences across the conditions (Table 5). To test our model-derived hypotheses, we calculated the probability of correction for the initially incorrect trials, excluding four participants who did not make any initial errors, which still left us with sufficient power to test our hypotheses ($N = 1{,}093$).

We found that, on average, the participants' probability of correcting their initial errors was relatively modest (one-third) and varied across conditions (see Table 5). The participants with the increased cost associated with the correction were slightly less likely to correct their errors, whereas those rewarded with a performance-based payment were more likely to correct their errors relative to the baseline condition (see Figure 5). The omnibus ANOVA testing for these differences was statistically significant, $F(2,\ 1{,}090) = 6.47$, $p = .002$, Cohen's $f = 0.11$, and it was followed by two planned contrasts to test our focal hypotheses. First, we tested the effect of both terms relative to the baseline. To do so, we reversed the effect of cost relative to the baseline mean to avoid the effect of cost and reward canceling each other out. The effect of cost (reversed relative to the baseline mean) and reward statistically significantly increased the probability of correction relative to the baseline, with a small effect size, $F(1,\ 1{,}090) = 4.27$, $p = .039$, Cohen's $f = 0.06$. Second, we tested the differential effect of the two terms. As predicted, the effect of cost and reward statistically significantly changed the probability of correction: Cost decreased the probability of correction, whereas reward increased the probability, with a small effect size, $F(1,\ 1{,}090) = 12.88$, $p < .001$, Cohen's $f = 0.13$. Thus, we found support for both hypotheses—the negative effect of cost and positive effect of reward—on the probability of correction (Hypothesis e4.1) and their differential effect (Hypothesis e4.2).

**Table 5**
*Effect of Cost and Reward on Reasoning and Counting Accuracy*

| Accuracy variables | Baseline | Cost | Reward | Total |
|---|---|---|---|---|
| | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) |
| All trials | | | | |
| Correct responses (%) | | | | |
| First stage | 26.3 (23.5) | 28.6 (21.9) | 27.5 (23.9) | 27.7 (23.5) |
| Incorrect valid trials | | | | |
| *P*(correction) | .35 (.29) | .31 (.27) | .39 (.32) | .35 (.30) |
| Counting accuracy | 0.38 (0.31) | 0.30 (0.31) | 0.53 (0.32) | 0.41 (0.33) |

*Note.* All trials: $N = 1{,}097$; incorrect valid trials: $N = 1{,}093$ ($n$ per condition in the order listed in the table: $n = 363$, $n = 366$, $n = 364$). The cost and reward conditions are relative to the task: For reasoning accuracy, reward refers to the condition in which participants were rewarded for reasoning accuracy, whereas for counting accuracy, reward refers to the condition in which participants were rewarded for the counting task.

### *Effect of Cost and Reward on Counting Accuracy*

We were also interested in the effect of cost and reward on accuracy in the counting task. On average, participants achieved around 41% accuracy with some remarkable differences across the conditions (Table 5). To aid interpretation in this section, we refer to cost and reward relative to the counting task (e.g., reward here refers to the condition in which participants were rewarded for the counting task). The participants with the increased cost for completing the counting task were less accurate, whereas those rewarded with a performance-based payment for the counting task were more accurate relative to the baseline condition (see Figure 5). We used identical analytical steps to test our hypotheses as above on the same sample ($N = 1{,}093$).

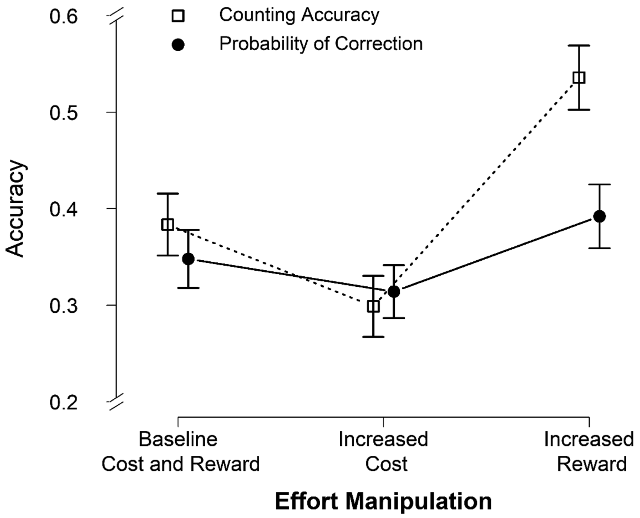
The omnibus ANOVA was statistically significant, $F(2, 1{,}090) = 53.25$, $p < .001$, Cohen's $f = 0.31$. We conducted two planned contrasts to test our focal hypotheses. First, the effect of the cost (reversed relative to the baseline mean) and reward statistically significantly increased counting accuracy relative to the baseline, with a medium effect size, $F(1, 1{,}090) = 34.48$, $p < .001$, Cohen's $f = 0.18$. Second, we tested the differential effect of the cost and reward, statistically significantly affecting accuracy, with a large effect size, $F(1, 1{,}090) = 103.65$, $p < .001$, Cohen's $f = 0.38$. Thus, both hypotheses —the effect of cost and reward on counting accuracy (Hypothesis e4.3) and their differential effect (Hypothesis e4.4)—were confirmed.

## Discussion

The results of Experiment 4 further supported the critical predictions of the EVC model. The cost and reward affected the probability of correction in reasoning over cognitive reflection problems in the expected direction. Two important points are worth mentioning. First, the observed effects of cost and reward were more substantial for the counting accuracy. This might be because of the varying complexity of the tasks. The cognitive reflection problems require complex problem representation skills and an ability to execute mathematical/verbal reasoning operations. On the other hand, counting tasks require much simpler counting skills, which we can reasonably expect in all our participants. Thus, intensifying effort may produce more impressive outputs in counting tasks than in problem-solving tasks. Second, our manipulation clearly varies cost as opportunity cost, which might be indexing the cost and benefit computations of the deployment of computational mechanisms (Kurzban et al., 2013); however, it might not necessarily vary the intrinsic cost of control (e.g., Kool & Botvinick, 2013). The manipulation might vary the intrinsic cost indirectly; for instance, there is an intrinsic cost in choosing between the two competing tasks, in trying to perform the two tasks simultaneously, in overcoming the possible temptation to start with the easier task or rushing through the harder task. Nevertheless, future research should manipulate the intrinsic cost directly.

**Figure 5**

*Effects of Cost and Reward on Counting and Reasoning Accuracy for an Originally Incorrect Response (i.e., the Probability of Correction)*



*Note.* $N = 1{,}093$; circles and squares represent means; error bars represent 95% confidence intervals. The probability of correction and counting accuracy ranges from 0 to 1. The cost and reward conditions are relative to the task; for instance, for reasoning accuracy, reward refers to the condition in which participants were rewarded for reasoning accuracy, whereas for counting accuracy, reward refers to the condition in which participants were rewarded for the counting task.

## Experiment 5

In Experiment 5, we extended the study of control allocation over corrective reasoning while making the step of the correction duration explicit. Specifically, in this experiment, participants first tried to solve a cognitive reflection problem within a restrictive time limit. Then, those who did not solve it correctly could decide how much control over reasoning to allocate in the correction stage. We measured the participants' decisions by asking whether they wanted to correct their answer and how much time they wanted to allocate to the problem. This was in order for us to demonstrate the effect of reward directly on the control allocation measure and, the downstream effect, accuracy (i.e., the probability of correction). Aligned with the EVC model, we expected that the reward would increase the amount of time people allocated to making their corrections (Hypothesis e5.1) and the probability of correcting their reasoning errors (Hypothesis e5.2) relative to the baseline.

## Method

### *Participants*

We recruited 401 participants from the online panel Prolific in exchange for a flat fee of £0.40 for a 4-min questionnaire. The sample size was based on an a priori stopping rule to reach the sample size of 400 participants; the additional participant was due to the online recruiting method implemented by Prolific. The target sample size was based on a power analysis. To detect a small effect of Cohen's $d = 0.4$ in an independent samples $t$-test $\alpha = 0.05$ and $\beta = 0.20$, we needed $N = 200$ participants. However, based on a pretest, we assumed around 50% of participants would answer correctly and not have a reason for corrective reasoning, so we doubled the final number. Participants were eligible to take part only if they had a minimal 80% approval rate in previous studies, had not participated in the previous studies using the CRT run by our lab, were UK nationals, and were at least 18 years old. None of the participants was excluded.

We asked 401 participants (50.1% women, 49.6% men, and 0.2% of another gender; with ages ranging from 18 to 79 years, $M = 38.2$, $SD = 12.6$ years) on Prolific to complete an online questionnaire using a sex-balanced sampling strategy. The sample varied in terms of education: 1.7% had less than a high school education, 33.7% had completed high school, 43.9% had a college degree, 17.2% had a master's degree, and 3.5% had a PhD or other professional degree. The sample also varied in terms of occupation; the most common occupation categories were management and professionals (33.4%), followed by government workers (8.7%), sales and office workers (8.0%), students (7.4%), unemployed (6.5%), and some other less common categories.

### Design

We used yet another modification of the two-response paradigm (Thompson et al., 2011). The participants first answered a cognitive reflection problem within a restrictive time limit of 15 s. We told participants whether their response was correct or not. Then, those who did not answer or answered incorrectly were offered an option to correct their answer and choose the time they would allocate to it (up to 60 s). At this stage, they were offered either no additional reward in the control condition or an extra payment of £0.30 for the correct answer in the reward condition. The experiment was a double-blind, randomized, controlled trial with automatic randomization of the conditions.

### Materials and Procedure

After providing informed consent, participants answered some tasks unrelated to this experiment and then proceeded to the task reported here. Participants were instructed to solve a cognitive reflection problem within a restrictive time limit of 15 s to elicit an intuitive response. This was the adapted version of the Mary's father problem used in Experiment 4, which prevented participants from easily finding the solution on the internet. (Mark's father loves ales, politics, and the United Kingdom. He has four children, and the names of the first three are England, Wales, and Scotland. The name of the fourth child is: …) Participants could choose from four answer options. After the time limit, they were provided with answer feedback (i.e., whether their answer was correct or not). Those whose answer was not correct were offered an opportunity to correct their response or to continue the study without correction. ("Do you want to skip to the next part or correct your response?": "Skip this problem," "Correct your answer.") If participants selected the correction option, they chose how much time to spend on the problem by selecting one of six time periods. ("You can choose now for how long you will be able to see and answer the problem. How much time do you want to spend on the problem?": "10 s," "20 s," … "60 s.") Critically, when choosing whether to correct their answer or not, those in the reward condition also learnt that they would get an extra payment for correctly answering the problem. ("We will pay you an extra £0.30 as a bonus payment if you answer the problem correctly.") The problem was then presented for the selected period of time (e.g., 30 s). Finally, participants completed a short questionnaire unrelated to this research, answered a sociodemographic question and were debriefed.

## Results

In the first stage of answering, 44.4% of participants correctly solved the problem, 44.9% answered incorrectly, and 10.7% did not answer within the time limit. This means that a large number of participants ($n = 223$) did not answer the problem correctly, which provided sufficient power to test our hypotheses.

Most participants attempted to correct their answers (95.1%), although the allocated time ($M = 29.4$ s, $SD = 18.7$ s) was far from the maximum time they could have allocated for the correction. To test our hypotheses, we treated the decision to skip the problem as a willingness to invest zero seconds in the correction and combined it with the measure of willingness to invest from 10 to 60 s in the correction. We observed variability across the conditions. On average, those in the reward condition tended to be willing to invest more time ($M = 32.4$, $SD = 19.1$, $n = 116$) than those in the baseline condition ($M = 26.1$, $SD = 17.8$, $n = 107$). This difference was statistically significant and of a small-to-medium effect size, $t(220.99) = -2.56$, $p = .011$, Cohen's $d = 0.34$. Thus, we found support for the hypothesis that a reward would increase the time allocated to the correction.

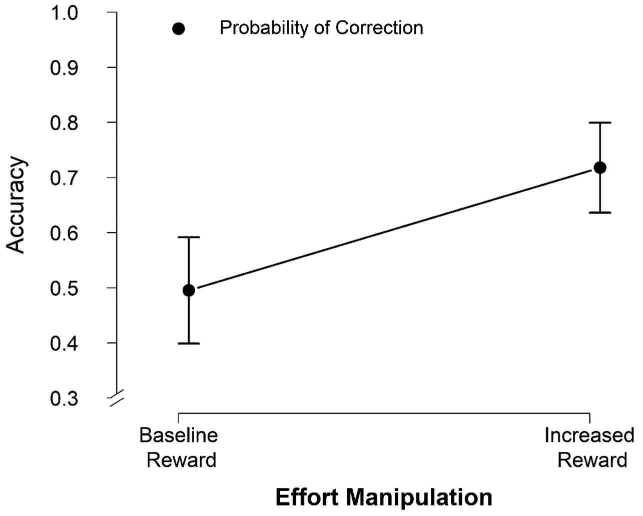
Regarding accuracy, most participants (63.0%) who attempted to correct themselves corrected their initial error. We also observed substantial variability across the conditions. Those in the reward condition were more likely to correct their mistake than those in the baseline condition (72.6% vs. 49.5%, respectively; see Figure 6). This difference was statistically significant and of a medium effect size, $\chi^2(1) = 10.94$, $p < .001$, Cramer's $V = 0.23$. Thus, we confirmed our hypothesis that the reward would increase the probability of correction compared with the baseline condition.

## Discussion

In Experiment 5, we further confirmed the critical predictions of the cognitive control literature over control allocated to corrective reasoning in a verbal cognitive reflection problem. The reward increased the time participants allocated for correction.

**Figure 6**

*Effect of Reward on Reasoning Accuracy for an Originally Incorrect Response (i.e., the Probability of Correction)*



*Note.* $N = 212$; circles represent estimated sample proportions; error bars represent 95% Agresti-Coull add-4 confidence intervals for a binomial proportion. The probability of correction ranges from 0 to 1.

Importantly, it also substantially boosted the probability of correction (i.e., accuracy) relative to the baseline condition, in which there was no additional financial reward given for performance.

## General Discussion

Do people always correct their errors when they know they have committed them? It is often assumed that if they do not, it is only because they are unable to. In this paper, we investigated the possibility that people might choose not to correct their errors even when they know how to. In five well-powered experiments, we studied whether and when people correct their reasoning errors when they are *unexpectedly* exposed to the same cognitive reflection problems a second time. In Experiment 1, we found that after repeated exposure to the same open-ended problems, participants spontaneously corrected some of their errors. Participants were more likely to correct their errors when they received answer feedback and when they were rewarded for the correct responses. In Experiments 2 and 3, participants virtually always corrected their errors in two-answer-choice problems when provided with explanatory feedback. Critically, however, increased cognitive costs imposed by an additional categorization task associated with the correction systematically depressed the probability of correction (Experiments 2 and 3), which was counteracted by increasing the reward associated with the correction (Experiment 3). In Experiment 4, cost decreased, and reward increased the probability of correction. Finally, in Experiment 5, reward increased the time that participants allocated to correction and the probability of correction (i.e., accuracy). We found reward and cost had medium-sized effects on correction decision-making underpinning the allocation of control to an additional task required to proceed with the correction (Experiments 2 and 3) and small- to medium-sized effects on control over corrective reasoning (Experiments 1, 4, and 5). We also found that participants were aware of the reasons for non-correction and that the feedback, cost, and reward effects were mostly manifested as related confidence changes.

These findings corroborate the idea that people perform a cost and benefit analysis before engaging in cognitive effort allocation. This supports and complements models developed to address motivational aspects of cognitive control using cognitively simpler tasks (Shenhav et al., 2014, 2016, 2017). For instance, according to the EVC model, reasoners monitor their performance (e.g., negative feedback, internal conflicts) and adjust how much and what kind of control to engage in. To do so, they evaluate their perceived efficacy to correct the error while taking into consideration the costs and benefits associated with the correction. The evidence we present supports these ideas in several respects. First, repeated exposure to the same problem allowed participants to allocate more mental effort to the same problem, which manifested in higher levels of spontaneous correction. In prior research, when participants were asked to second-guess their answer to a bat-and-ball problem while excluding their original error, they selected the option closer to the correct response but did not provide a higher rate of correct answers (Bago et al., 2019). Here, however, we had higher power aggregated across many problems and participants to detect such a possible effect. Second, the positive effect of answer feedback we found in Experiment 1 on the probability of correction can enhance the monitoring process and signal to a reasoner that more/a different kind of control is needed. This is aligned with prior reasoning research which found that feedback prolonged the processing time and, in turn, improved reasoning, at least for some participants (Ball, 2013; Ball et al., 2010; Janssen et al., 2020). The effect of explanatory feedback is also aligned with these findings even though, admittedly, the effect is trivial because the participants also learned the correct response in the feedback. The purpose of such feedback was to maximize the control efficacy of our participants, so we could observe the effects of cost and reward independently of efficacy. Finally, the positive effect of reward and the negative effect of cost on the probability of correction shows that control allocation follows the cost-and-benefit analysis (Shenhav et al., 2014). As such, this evidence is aligned with the wider cognitive psychology literature, which demonstrates the positive effect of external reward on performance (e.g., Camerer & Hogarth, 1999; Neyse et al., 2016), as well as correlational evidence for the association between cognitive cost and internal rewards of control and performance (Shenhav et al., 2017).

Our findings also have implications for the existing models of dual processes that currently dominate research on reasoning and judgment. First, our findings further demonstrate the theoretical need to untangle error detection and error correction processes in dual-process theories. Some current dual-process theories implicitly assume that once the reasoning error is detected, people will correct it automatically (e.g., Kahneman & Frederick, 2002). This assumption was criticized on theoretical grounds and empirically rebutted in the domain of belief formation such as superstitious beliefs (Risen, 2016, 2017; Walco & Risen, 2017). Here, we complemented such evidence by demonstrating that people do not always correct their reasoning errors, even when they have detected the error in their reasoning and they have sufficient knowledge available to correct it. Thus, decoupling error detection and correction processes in models of reasoning beyond belief formation can help us better understand how people reason strategically.

Second, the cognitive control literature extends the dual-process theories by specifying the motivational factors influencing the decision to engage in correction. So far, we know that motivation to engage in mentally effortful activity leads to improved reasoning (e.g., Stanovich & West, 1998), but the understanding of the factors underlying this motivation is rudimentary. The EVC model, for example, provides a useful framework to decompose motivational factors and study them systematically for both error detection and error correction processes. In future, dual-process theories should try to integrate motivational forces of error detection and correction explicitly in their models. Approximating such an extension, Evans (2011) proposed that motivational factors play an important role in setting the level of critical effort, which determines correction processes. The current model can extend this by decomposing the motivational factors into perceived efficacy in detecting and correcting errors as well as benefits and costs associated with such processes. Engagement of Type II processes during the error correction process, specifically with cognitive reflection problems, can result in better text comprehension and more appropriate problem representation, carrying out calculations more carefully and double-checking answers, or deploying more effective problem-solving strategies.

Third, in a more speculative endeavor, our findings might shed more light on the underlying motivational mechanism of two different Type II processes: correction and justification. Some dual-process models formalized (Evans, 2011; Pennycook et al., 2015) or at least thematized (Bago & De Neys, 2020; De Neys &

Pennycook, 2019) two different Type II processes as (a) correcting the default answers and (b) justification/rationalization of the default answers. These models remained mostly silent about the mechanism deciding whether the correction or justification should be triggered. However, Evans (2011) proposed that motivational factors play an important role in setting the level of critical effort, which determines whether a reasoner will endorse the default answer as justified or try to correct it; thus, the justification of the default answer happens first. In line with this proposal, we speculate here that a reasoner can calculate the EVC needed for justification of the default answer and correction of the default answer if recognized as wrong. For instance, a reasoner might unsuccessfully attempt to correct a default incorrect answer, dynamically lowering her estimate of the probability of reaching the correct answer and, in turn, switch into the justification of the default answer. Future research might test whether such considerations take place and which components affect such decision-making more profoundly.

As with any research, our study features many limitations that should be addressed in future work; three of them should be discussed in detail here. First, though not consequential to our conclusions, our manipulation of cost and reward can be improved. To directly manipulate the correction costs in our research, we "externalized" the mental effort and associated it with a secondary task (Experiments 2 and 3). Such an approach has some advantages (e.g., controlling for efficacy) but also limits us to imposing control over an alternate task and not the reasoning task itself. Future studies assessing the effects of cost on control over reasoning could ensure those cost manipulations are intrinsic to reasoning and affect control over reasoning more directly. In Experiment 4, we offered one possible manipulation of cost affecting control over reasoning. However, future research should finetune such operationalizations and also explore other ways to measure and manipulate costs directly. This might be achieved, for instance, by manipulating the perception of the expected cost for each problem or using the "externalized" correction cost but with open-ended cognitive reflection problems or structurally identical cognitive reflection problems with a modified content to make sure reasoning is taking place.

Similarly, we focused on the external reward, which was, in addition, relatively small in comparison with previous research (e.g., Frömer et al., 2021; Neyse et al., 2016). In our experiments, the reward was always contingent on the performance (the correct answer) even though the monetary value varied across the experiments (Experiments 1, 3, 4, and 5). This effect of different monetary values might be an interesting empirical question for future studies. Other operationalization of rewards such as enhanced social reputation and accountability might also be important to test. Furthermore, intrinsic rewards such as high satisfaction of solving a problem might be tricky to manipulate but worthy of exploring in future research since these might be the main source of reward in everyday life for people solving tricky word problems correctly.

Second, we focused on explicit error detection using external feedback. We assumed that similar mechanisms operate over internal conflict monitoring. Future research should test the validity of this assumption empirically. Finally, we focused on feedback and how this informs the conditional probability of correction given the exerted effort. However, the explanatory feedback confounds perceived efficacy with providing knowledge. The perceived control efficacy should be manipulated independently in future research given the important role that perceived efficacy plays in reasoning, for instance, the role of mathematical anxiety in solving numerical cognitive reflection problems (e.g., Juanchich et al., 2020). Researchers could try to manipulate the perceived problem-solving and control efficacy directly by manipulating the contingency between control efficacy and reward (Frömer et al., 2021).

To conclude, people do not automatically correct the errors they have unequivocally detected even when they have sufficient knowledge to do so. Instead, they decide whether or not to correct these errors following rational principles based on the value of control optimization. Paradoxically, following such principles might lead to keeping more reasoning errors, dubbed as irrational, in the cognitive reflection reasoning problems. The presented evidence emphasizes the importance of motivational factors in reasoning. It corroborates the EVC model applied to reasoning and extends the currently dominating models of dual-process theories.

## References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Bago, B., Bonnefon, J.-F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*, *150*(6), 1081–1094. https://doi.org/10.1037/xge0000968

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

Bago, B., & De Neys, W. (2019a). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801. https://doi.org/10.1037/xge0000533

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

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bago, B., Raoelison, M., & De Neys, W. (2019). Second-guess: Testing the specificity of error detection in the bat-and-ball problem. *Acta Psychologica*, *193*, 214–228. https://doi.org/10.1016/j.actpsy.2019.01.008

Ball, L. J. (2013). Microgenetic evidence for the beneficial effects of feedback and practice on belief bias. *Journal of Cognitive Psychology*, *25*(2), 183–191. https://doi.org/10.1080/20445911.2013.765856

Ball, L. J., Hoyle, A. M., & Towse, A. S. (2010). The facilitatory effect of negative feedback on the emergence of analogical reasoning abilities. *British Journal of Developmental Psychology*, *28*(3), 583–602. https://doi.org/10.1348/026151009X461744

Banks, A. P., & Hope, C. (2014). Heuristic and analytic processes in reasoning: An event-related potential study of belief bias. *Psychophysiology*, *51*(3), 290–297. https://doi.org/10.1111/psyp.12169

Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (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. https://doi.org/10.1016/j.jarmac.2014.09.003

Bhatia, S. (2017). Conflict and bias in heuristic judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(2), 319–325. https://doi.org/10.1037/xlm0000307

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital–labor–production framework. *Journal of Risk and Uncertainty*, *19*(1), 7–42. https://doi.org/10.1023/A:1007850605129

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cooper-Martin, E. (1994). Measures of cognitive effort. *Marketing Letters*, *5*(1), 43–56. https://doi.org/10.1007/BF00993957

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, *17*(5), 428–433. https://doi.org/10.1111/j.1467-9280.2006.01723.x

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, *6*(1), Article e15954. https://doi.org/10.1371/journal.pone.0015954

De Neys, W., & Feremans, V. (2013). Development of heuristic bias detection in elementary school. *Developmental Psychology*, *49*(2), 258–269. https://doi.org/10.1037/a0028320

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

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*(2), 269–273. https://doi.org/10.3758/s13423-013-0384-5

Dunn, T. L., Lutes, D. J., & Risko, E. F. (2016). Metacognitive evaluation in the avoidance of demand. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1372–1387. https://doi.org/10.1037/xhp0000236

Enke, B., Gneezy, U., Hall, B., Martin, D. C., Nelidov, V., Offerman, T., & van de Ven, J. (2021). Cognitive biases: Mistakes or missing stakes? *The Review of Economics and Statistics*, 1–45. https://doi.org/10.1162/rest_a_01093

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, *31*(2–3), 86–102. https://doi.org/10.1016/j.dr.2011.07.007

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

Evans, J. S. B. T., & Over, D. E. (2013). *Rationality and reasoning*. Psychology Press.

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

Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, *12*(1), Article 1030. https://doi.org/10.1038/s41467-021-21315-z

Healy, A. F., & Nairne, J. S. (1985). Short-term memory processes in counting. *Cognitive Psychology*, *17*(4), 417–444. https://doi.org/10.1016/0010-0285(85)90015-5

Heled, E., Hoofien, D., Margalit, D., Natovich, R., & Agranov, E. (2012). The Delis–Kaplan executive function system sorting test as an evaluative tool for executive functions after severe traumatic brain injury: A comparative study. *Journal of Clinical and Experimental Neuropsychology*, *34*(2), 151–159. https://doi.org/10.1080/13803395.2011.625351

Isler, O., Yılmaz, O., & Doğruyol, B. (2020). Activating reflective thinking with decision justification and debiasing training. *Judgment and Decision Making*, *15*(6), 926–938. https://doi.org/10.1017/S1930297500008147

Janssen, E. M., Raoelison, M., & de Neys, W. (2020). "You're wrong!": The impact of accuracy feedback on the bat-and-ball problem. *Acta Psychologica*, *206*, Article 103042. https://doi.org/10.1016/j.actpsy.2020.103042

Juanchich, M., Sirota, M., & Bonnefon, J.-F. (2020). Anxiety-induced miscalculations, more than differential inhibition of intuition, explain the gender gap in cognitive reflection. *Journal of Behavioral Decision Making*, *33*(4), 427–443. https://doi.org/10.1002/bdm.2165

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

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). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press.

Kool, W., & Botvinick, M. (2013). The intrinsic cost of cognitive control. *Behavioral and Brain Sciences*, *36*(6), 697–698. https://doi.org/10.1017/S0140525X1300109X

Kool, W., Shenhav, A., & Botvinick, M. M. (2017). Cognitive control as cost-benefit decision making. In T. Egner (Ed.), *The Wiley handbook of cognitive control* (pp. 167–189). Wiley.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679. https://doi.org/10.1017/S0140525X12003196

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), 660–684. https://doi.org/10.1017/S1930297500007865

Logie, R. H., & Baddeley, A. D. (1987). Cognitive processes in counting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(2), 310–326. https://doi.org/10.1037/0278-7393.13.2.310

Mamede, S., Schmidt, H. G., & Rikers, R. (2007). Diagnostic errors and reflective practice in medicine. *Journal of Evaluation in Clinical Practice*, *13*(1), 138–145. https://doi.org/10.1111/j.1365-2753.2006.00638.x

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

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *8*(1), 3–30. https://doi.org/10.1145/272991.272995

Neyse, L., Bosworth, S., Ring, P., & Schmidt, U. (2016). Overconfidence, incentives and digit ratio. *Scientific Reports*, *6*(1), Article 23294. https://doi.org/10.1038/srep23294

Patel, N., Baker, S. G., & Scherer, L. D. (2019). Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs. *Journal of Experimental Psychology: General*, *148*(12), 2129–2153. https://doi.org/10.1037/xge0000592

Pennycook, G. (2018). A perspective on the theoretical foundation of dual process models. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 5–27). Routledge/Taylor & Francis Group.

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

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction*, *12*(1), 61–86. https://doi.org/10.1016/S0959-4752(01)00016-0

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

Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review*, *123*(2), 182–207. https://doi.org/10.1037/rev0000017

Risen, J. L. (2017). Acquiescing to intuition: Believing what we know isn't so. *Social and Personality Psychology Compass*, *11*(11), Article e12358. https://doi.org/10.1111/spc3.12358

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, *53*(1), 491–517. https://doi.org/10.1146/annurev.psych.53.100901.135213

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

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience*, *19*(10), 1286–1291. https://doi.org/10.1038/nn.4384

Shenhav, A., Fahey, M. P., & Grahek, I. (2021). Decomposing the motivation to exert mental effort. *Current Direction in Psychological Science*, *30*(4), 307–314. https://doi.org/10.1177/09637214211009510

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*(1), 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Neuroscience*, *17*(9), 1249–1254. https://doi.org/10.1038/nn.3771

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

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., & Marshall, A. C. (2021). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making*, *34*(3), 322–343. https://doi.org/10.1002/bdm.2213

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the cognitive reflection test. *Behavior Research Methods*, *50*(6), 2511–2522. https://doi.org/10.3758/s13428-018-1029-4

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic Bulletin & Review*, *21*(1), 198–204. https://doi.org/10.3758/s13423-013-0464-6

Sirota, M., Juanchich, M., & Holford, D. L. (2022). *Rationally irrational: When people do not correct their reasoning errors even if they could*. OSF. https://doi.org/10.17605/OSF.IO/HVQA4

Sirota, M., Theodoropoulou, A., & Juanchich, M. (2021). Disfluent fonts do not help people to solve math and non-math problems regardless of their numeracy. *Thinking & Reasoning*, *27*(1), 142–159. https://doi.org/10.1080/13546783.2020.1759689

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Srol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning*, *27*(1), 38–68. https://doi.org/10.1080/13546783.2019.1708793

Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Psychology Press.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. https://doi.org/10.1037/0096-3445.127.2.161

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

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the cognitive reflection test. *Cognition*, *150*, 109–118. https://doi.org/10.1016/j.cognition.2016.01.015

Walco, D. K., & Risen, J. L. (2017). The empirical case for acquiescing to intuition. *Psychological Science*, *28*(12), 1807–1820. https://doi.org/10.1177/0956797617723377

Received September 16, 2021
Revision received November 3, 2022
Accepted January 10, 2023 ■