Moses illusions, fast and slow

Jérémie Beucler[a,*], Aikaterini Voudouri[a], Wim De Neys[a]

[a] Université Paris-Cité, LaPsyDÉ, CNRS, F-75005 Paris, France ; jeremie.beucler@gmail.com (J. Beucler); aikaterini.voudouri@gmail.com (A.Voudouri); wim.de-neys@parisdescartes.fr (W. De Neys)

*Corresponding author at: Université Paris-Cité, LaPsyDÉ, CNRS, 46, rue Saint-Jacques, F-75005 Paris, France. E-mail address: jeremie.beucler@gmail.com (J. Beucler).

Abstract

When asked “How many animals of each kind did Moses take on the Ark?”, most people answer “Two”, failing to notice that it was Noah, and not Moses, who took the animals in the Ark. “Fast-and-slow” dual process accounts of such semantic illusions posit that incorrect responders are not sensitive to their error and that overcoming the illusion requires deliberate correction of an intuitive erroneous answer. We present three experiments that force us to revise this dual process view. We used a two-response paradigm in which participants had to give their first, initial answer under cognitive load and time pressure. Next, participants could take all the time they wanted to deliberate and select a final answer. This enabled us to identify the intuitively generated response that preceded the final response given after deliberation. Results show that participants do not necessarily need to deliberate to avoid the illusion and that incorrect respondents consistently display error sensitivity (as reflected in decreased confidence), even when deliberation is minimized. Both reasoning performance and error sensitivity in the initial, intuitive stage tended to be driven by the semantic relatedness between the anomalous word (e.g., “Moses”) and the undistorted word (e.g., “Noah”). We show how this leads to a revised model where the response to semantic illusions depends on the interplay of both incorrect and correct intuitions.

Keywords: semantic illusions; Moses illusion; dual process theory; two response paradigm

Moses Illusions, Fast and Slow

When asked "How many animals of each kind did Moses take on the Ark?", most people answer "Two", failing to notice that it was Noah, and not Moses, who took the animals in the Ark (Erickson & Mattson, 1981). This tendency to overlook a semantic anomaly in a sentence is known as a *semantic illusion* (or the *Moses illusion*—after its most famous example). It is a very robust effect that attracted a lot of attention in the memory field and beyond (e.g., Cantor & Marsh, 2017; Hannon & Daneman, 2001; Kamas et al., 1996; Park & Reder, 2004; Reder & Kusbit, 1991; Shafto & MacKay, 2000; Speckmann & Unkelbach, 2021).

A key driving factor in the emergence of the illusion is the semantic relatedness between the anomalous (or distorted) word (e.g., "Moses") and the correct or undistorted word (e.g., "Noah," Erickson & Mattson, 1981; Hannon & Daneman, 2001; Van Oostendorp & De Mul, 1990). In particular, Noah and Moses share many semantic attributes such as biblical character, male, leader, and so on. The anomalous word will thus serve as an "impostor" and go unnoticed because of how semantically similar it is to "Noah". Indeed, if the semantic similarity between the distorted and undistorted word is low (e.g., "How many animals of each kind did Nixon take on the Ark?"), participants are much less likely to fall prey to the illusion and will respond correctly that the question is anomalous and cannot be answered (e.g., Erickson & Mattson, 1981; Hannon & Daneman, 2001; Van Oostendorp & De Mul, 1990).

More generally, semantic illusions provide a relevant testbed for examining the human tendency for miserly processing or satisficing such as it has been put forward in the dual process framework (Kahneman, 2011; Koriat, 2017; Stanovich & West, 2000). This influential framework conceives human cognition as an interplay of fast and effortless, intuitive (“System 1”) processing, and slower, more effortful deliberate (“System 2”) processing (Evans & Stanovich, 2013; Kahneman, 2011). Although fast intuitive processing may often cue valid responses, it can sometimes also cue responses that conflict with the slower, deliberate processing and will need to be corrected. However, because people will typically try to minimize spending cognitive effort, they will often refrain from engaging the effortful processing. Consequently, they will fail to detect that their intuitively cued response is erroneous and end up with a biased judgment (Evans & Stanovich, 2013; Kahneman, 2011).

In language comprehension, the “Good-Enough” processing framework also distinguishes between a fast, heuristic processing and a slower, algorithmic processing (e.g., Christianson, 2016; Ferreira & Huettig, 2023; Ferreira & Patson, 2007). At its core, this approach suggests that language comprehension operates in parallel through two distinct modes: a heuristic processing using fast and frugal heuristics, and a deeper algorithmic processing. Due to its rapidity, the heuristic processing, though shallower, often outpaces the algorithmic processing, potentially resulting in incorrect representations (Karimi & Ferreira, 2016).

In the case of semantic illusions, Park and Reder (2004) argued that people rely on an automatic partial matching mechanism that focuses on the coarse fit between a memory trace and the presented sentence. As long as there is sufficient semantic overlap, people will not engage in a more effortful in-depth analysis. Although the fast-matching mechanism might often be useful for quick sentence comprehension, it can also give rise to semantic illusions. Hence, semantic illusions may allow us to generalize the study of a more general human failure to switch from fast intuitive to slow deliberate processing when it is needed (Koriat, 2017; Mata et al., 2013). For example, semantic illusions feature in widely used Cognitive Reflection Tests that are intended to measure people’s capacity and disposition to engage in effortful deliberation rather than to stick to a mere intuitive hunch (e.g., Sirota et al., 2021).

At first sight, the dual process account of semantic illusions does not seem unreasonable. From an introspective point of view, many of us will have fallen for the illusion and attest that spotting it requires taking more time for deeper reflection and having a closer, second look. There is also some empirical evidence that is consistent with the account. For example, several lines of research have demonstrated the importance of cognitive resources to avoid the illusion. Experiments using a concurrent cognitive load have shown that burdening participants’ cognitive resources increases the likelihood of falling prey to the illusion (Büttner, 2012; Mata et al., 2013). In addition, working memory capacity is positively correlated with the ability to detect the illusion, suggesting that cognitive resources are needed to overcome it (Hannon & Daneman, 2001).

Nevertheless, the available evidence does not tell us whether deliberation is always necessary to correctly detect the anomaly and avoid the illusion. In theory, it is possible that in addition to the slow route there is also a fast route to anomaly detection in which the correct answer is cued intuitively. That is, rather than correcting an incorrect hunch (i.e., “Two”) after having taken the time to deliberate, people might generate the correct answer from the outset. Clearly, if one intuitively detects the anomaly and avoids the illusion, there is no further need to deliberately correct it. Obviously, such a fast route would imply that it would be problematic to use people’s performance on semantic illusion items as a measure of cognitive reflection or deliberation.

Recent dual process studies in the reasoning field lend some credence to this theoretical possibility (e.g., see De Neys & Pennycook, 2019). In these studies, participants solve logico-mathematical “bias” problems in which intuitive processing can lead them astray (e.g., the notorious bat-and-ball problem, “A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?"; correct answer: "5 cents", incorrect intuitive answer: "10 cents"). To isolate more intuitively and deliberately generated responses the studies use a two-response paradigm (Thompson et al., 2011), in which participants have to initially give the first response that comes to mind as quickly as possible and are subsequently given all the time they want to deliberate and select a final response. To be maximally sure that participants do not engage in deliberation in the initial stage, they are forced to give their first response under time-pressure while performing a concurrent cognitive load task which burdens their cognitive resources (Bago & De Neys, 2017). Since deliberation takes time and cognitive resources, depriving people from these resources minimizes the possibility that participants will deliberate before giving their initial response. The traditional corrective dual process view predicts that correct responses will only emerge after deliberation in the final response stage (Kahneman, 2011). However, contrary to this view, results across a range of reasoning problems have now shown that on those trials where participants manage to give the correct response after deliberation, they often already generate the same correct response during the initial response stage (e.g., Bago & De Neys, 2017, 2019a; Burič & Konrádová, 2021; Burič & Šrol, 2020; Raoelison et al., 2020; Thompson & Johnson, 2014). Hence, sound reasoners are not necessarily good at deliberately correcting erroneous intuitions, but rather at accurate intuiting (Raoelison et al., 2020; Thompson et al., 2018).

Further evidence against the postulated role of deliberation during reasoning comes from intuitive error detection findings (e.g., De Neys & Pennycook, 2019). Just as with the dual process view on semantic illusions, it is traditionally assumed that detecting that an intuitively cued solution is incorrect, requires that people engage in effortful deliberation (Evans & Stanovich, 2013; Kahneman, 2011). Contrary to this assumption, however, it has been found that reasoners who fail to generate the correct response to logical bias problems often show some intuitive sensitivity to their error (e.g., De Neys et al., 2011; Stupple & Ball, 2008; Stupple et al., 2011; Voudouri et al., 2022; but see also Mata et al., 2014, 2017). For example, they show lower confidence in their erroneous responses than in their correct responses to control problems. Critically, this error sensitivity is also observed in the initial stage of the two-response studies (i.e., when deliberation is minimized with time-pressure and load, e.g., Bago & De Neys, 2017, 2019b; Bialek & De Neys, 2017; Burič & Konrádová, 2021; Burič & Šrol, 2020; Johnson et al., 2016; Pennycook et al., 2014; Thompson & Johnson, 2014). Hence, even in the absence of proper deliberation, participants seem to be able to detect that their erroneous answer is not fully warranted.

Taken together, recent findings in the logical reasoning field suggest that the traditional role of deliberation in the dual process framework can be questioned (De Neys, 2023; Evans, 2019; Stanovich, 2018). If we were to generalize this pattern to semantic illusions, this suggests that detecting the anomaly and avoiding falling prey to semantic illusions may also be done intuitively and might not necessarily require slow and effortful deliberation. This would have critical implications for our conceptualization of semantic illusions and their use as an index of deliberate processing abilities. In addition, since the dual-process framework is a general model of human cognition, we believe it is important to extend its revisions to phenomena which, although they are not logical reasoning tasks, have still been conceptualized using a dual process framework (e.g., Low et al., 2023; March et al., 2023).

To prevent any misunderstandings, we should note that we used the dual process framework and labels (“intuitive”; “deliberative”) because we believe they are a useful tool to communicate between scholars. However, in this study, we define “intuition” operationally. Intuitive processes were operationalized by combining instructions, time pressure, and concurrent load to minimize the engagement of cognitive resources in the initial response stage/block of our paradigm. Therefore, we do not take sides in the theoretical debate regarding whether there is a qualitative or quantitative difference between intuitive and deliberative reasoning processes (for an in-depth discussion, see De Neys, 2021). Instead, our focus was on generalizing recent dual-process findings in the reasoning field, departing from the traditional study of semantic illusions in the field of psycholinguistics.

In the present experiments we test the hypothesis that avoiding semantic illusions may be done intuitively directly by introducing a two-response paradigm of semantic illusions. Participants were presented with a range of trivia questions that are known to elicit semantic illusions. Half of the problems were presented in an undistorted format (e.g., "In the tale, who found the glass slipper left at the ball by *Cinderella*?") and served as control problems on which intuitive processing is expected to cue the correct response. The other half of the problems were classic "anomaly" problems (e.g., "In the tale, who found the glass slipper left at the ball by *Snow White*?") in which the undistorted word was replaced by a semantically related distorted "impostor" word which may give rise to a semantic illusion. Participants had to give an initial response as fast as possible (under time pressure and concurrent load) and immediately after could take the time to deliberate and give a final response. Participants also indicated their response confidence. Our key research questions were: First, whether people who answer anomaly problems correctly and avoid the illusion after deliberation, can also provide a correct response to these questions intuitively. Second, whether people who give an incorrect response to anomaly problems in the intuitive response stage, show error sensitivity (i.e., by contrasting their response confidence in the anomaly and control no-anomaly problems).

We present a set of three experiments: Experiment 1 introduces the paradigm and Experiment 2 tests the robustness of the results with methodological refinements. Finally, Experiment 3 introduces a direct manipulation of the semantic similarity factor, along with a two-block paradigm that contrasts an intuitive block with a deliberative block, to further expand and validate the findings.

Experiment 1

Method

Transparency and openness.

The research question and experimental design were preregistered on the OSF platform (https://osf.io/bpmc8). No specific analyses were preregistered. All data, material and analysis scripts can be retrieved from https://osf.io/bvy3u/.

Participants

In Experiment 1, we recruited 100 participants (78 females, *M* age = 32.6 years, *SD* = 12.2 years) on the Prolific platform (app.prolific.co). Only native English-speaking American (USA) participants were allowed to participate in the experiment. Among them, 42 reported high school as their highest level of education, 1 less than high school and 57 a higher education degree. Participants received £1.2 for their 12 minutes of participation.

We based our sample size decision for Experiment 1 and 2 on previous two-response work in the logical, moral, and economic reasoning field (Bago et al., 2021; Bago & De Neys, 2017, 2019a), which also tested approximately 100 participants. Note that this sample size gives us more power than most previous studies on semantic illusions (e.g., Kamas et al., 1996).

In addition, we conducted sensitivity power analyses to estimate our achieved power across a plausible range of effect sizes in our main analyses. These analyses evaluated whether our design had adequate power to detect various effect sizes during the hypothesis testing process (Lakens, 2022). Overall, the results indicated that we could detect small to medium effect sizes with more than 80% achieved power across our studies and analyses (see Supplementary Material A for the complete simulation results).

Materials

Trivia Questions. We selected 20 multiple-choice trivia question problems from the second experiment of Speckmann and Unkelbach (2021). In this experiment, 200 participants first answered 40 anomaly or (control) no-anomaly multiple-choice questions (e.g., “How many animals of each kind did Noah take on the Ark?”). Next, they responded to the corresponding 40 open-ended knowledge-check questions (e.g., “Which biblical figure took two animals of each kind on the Ark?”).

Our item selection was based on their knowledge check results, while ensuring that the selected item also had high enough control, no-anomaly accuracy in the actual multiple-choice experiment. This further helped to guarantee that on average our participants would know the correct answer to the original questions. Specifically, we selected the 20 questions that were above or closest to the knowledge check sample median accuracy (*Mdn* = 74.4%). We discarded one item with low control, no-anomaly accuracy. We also decided to discard one item because its high anomaly accuracy suggested it was too easy. We replaced these two items with the questions that were closest to the knowledge check median accuracy. We also introduced some superficial content modifications to minimize question length differences. Table 1 provides the complete list of anomaly and no-anomaly questions. The results from Speckmann and Unkelbach used to guide our selection are available on the OSF platform.

The resulting average knowledge-check accuracy in our selected pool of questions was 83.4%, which is by construction higher than in Speckmann and Unkelbach (2021), where the average knowledge accuracy was 69.7% in their additional open-ended norming study (n = 120) and 73.2% in Experiment 2. In addition, the consistently very high final accuracy on the undistorted no-anomaly questions in our own experiments (e.g., 92% in Experiment 1; 90.8% in Experiment 2) can be used as an additional control that our participants did know the undistorted answer to the selected questions. Still, we cannot exclude the possibility that some subjects may not know the correct response to specific items. In some studies, subjects are explicitly asked to generate answers to open-ended versions of the undistorted questions after the experiment, and the corresponding trial is excluded if the response to the open-ended version of the question is incorrect (e.g., Erickson & Mattson, 1981; Kamas et al., 1996). However, this conservative procedure has its limits, as participants' performance on this subsequent knowledge-check is likely to be affected by fatigue, as well as by the prior presentation of the distorted items. Indeed, the latter has been shown to decrease open-ended accuracy (Bottoms et al., 2010), thus challenging the validity of such a procedure. Nevertheless, Speckmann and Unkelbach (2021, Experiment 2), who used a less restrictive item set than ours, showed that excluding trials where participants failed the knowledge check had a very small impact on the illusion strength, reducing it from 54% to 52% in terms of the percentage of trials where participants fell prey to the illusion. Taken together, these results suggest that the impact of a potential lack of knowledge of the correct response to the undistorted question should be minimal in our experiments.

Table 1

Overview of the Stimuli Used in the Experiments

Question Number	No-Anomaly Question (Undistorted Answer; Filler Answer)	Strong Impostor	Weak Impostor
1	What kind of tree did the later president *Washington* allegedly chop down? (Cherry; Palm)	Lincoln	Nixon
2	In what movie did *Arnold Schwarzenegger* go back in time to protect Sarah Connor? (Terminator 2; Rocky 2)	Sylvester Stallone	Johnny Depp
3	What country was Margaret Thatcher *prime minister* of for several years? (United Kingdom; France)	President	Queen
4	In what year did Germany *lose* the second World War? (1945; 1918)	Win	Was the victor of
5	What kind of meat is in the *Burger King* sandwich known as the Whopper? (Beef; Chicken)	McDonald's	Taco Bell
6	What season do we associate with football games, starting school, and leaves turning *brown*? (Fall; Winter)	Green	Black
7	What statue given to the U.S. by *France* symbolizes freedom to immigrants arriving in New York? (Statue of Liberty; Christ the Redeemer)	England	Austria
8	Who is the video game character and Italian plumber who is *Nintendo*'s mascot? (Mario; Sonic)	Sony	Apple
9	In the tale, who found the glass slipper left at the ball by *Cinderella*? (The prince; The stepmother)	Snow White	Pocahontas
10	What is the name of the kimono-clad courtesans who entertain *Japanese* men? (Geisha; Samurai)	Chinese	French
11	Which instrument gives the *time* by measuring the angle of the sun's shadow on a dial? (Sundial; Oscillator)	Temperature	Humidity
12	What is the name of the comic strip character who eats spinach to improve his *strength*? (Popeye; Mickey Mouse)	Sight	Intelligence
13	What is the name of the current dictator of *North* Korea? (Kim Jong-Un; Fidel Castro)	South	East
14	What is the name of the molten rock coming out of a volcano during an *eruption*? (Lava; Mud)	Earthquake	Tsunami
15	How do we call the man in the red suit and white beard who gives *Christmas* presents from his sleigh? (Santa Claus; Rumpelstiltskin)	Birthday	Wedding
16	What is the name of the Mexican dip made with mashed-up *avocados*? (Guacamole; Salsa)	Artichokes	Cucumbers
17	What is the name of the scary carved pumpkin displayed on *Halloween*? (Jack-o'-lantern; Soul cake)	Thanksgiving	Easter
18	When did the *Japanese* attack Pearl Harbor with their planes during World War II? (December 7th, 1941; December 7th, 1951)	Germans	Vietnamese
19	What is the name of the New Year festival celebrated on the 31st of *December*? (New Year's Eve; Carnival)	January	March
20	In the biblical story, how many animals of each kind did *Noah* take on the Ark? (Two; Three)	Moses	Goliath

Note. The undistorted, original word in the no-anomaly question is italicized. The strong impostor has high semantic similarity, while the weak impostor (Experiment 3) has low semantic similarity to the original word.

For each of the selected questions we created an anomaly version (i.e., “In the biblical story, how many animals of each kind did Moses take on the Ark?”) and a control, no anomaly version (i.e., “In the biblical story, how many animals of each kind did Noah take on the Ark?”)[1]. The control version used the original, undistorted word (e.g., “Noah”) whereas the anomaly version used the semantically related “impostor” word (e.g., “Moses”) as in Speckmann and Unkelbach (2021). Half of the 20 problems that each participant saw were anomaly problems and the other half control problems. Two question sets were created for counterbalancing. For each question, the control version was used in one set and the anomaly version in the other set. Participants were randomly assigned to one of the sets. Hence, participants never saw the same question content more than once. The presentation order of the questions was randomized in both sets.

Following Speckmann and Unkelbach (2021), each question had four different response options. The first option was the “undistorted” answer (e.g., “two” for the Moses question) and could be correct or incorrect depending on the question version (no-anomaly vs. anomaly). The second option (e.g., “three”) was always incorrect. The third response option was “This question can’t be answered in this form” and could be correct or incorrect depending on the question version (anomaly vs. no-anomaly). The fourth option was “Don’t know”, which was always coded as incorrect. The order of options 1 and 2 was randomized, but we kept the order of response options 3 and 4 fixed so as not to confuse participants. Note that the use of a multiple-choice (vs. open-ended question) design with these specific response options was tested and validated across four experiments by Speckmann and Unkelbach (2021). Semantic illusions were as prevalent in the multiple-choice design as in previous open-ended studies. In addition, to be sure that participants

[1] Note that in our preregistration we referred to conflict and no-conflict problems (in analogy with the reasoning field). Here we have opted for the more descriptive anomaly and (control) no-anomaly labels.

understood the difference between the "Don't know" and the "This question can't be answered in this form" response options, the following examples were presented in the instructions:

> *What is the name of former president's Obama's oldest son?*
>
> *Charles*
> *Jonathan*
> *This question can't be answered in this form.*
> *Don't know.*
>
> The above question cannot be answered because Obama doesn't have a son; he only has two daughters. So, the correct answer option to this question is: 'This question can't be answered in this form.'
>
> Here is a different example:
>
> *What is the name of former president's Obama's oldest daughter?*
>
> *Sasha*
> *Malia*
> *This question can't be answered in this form.*
> *Don't know.*
>
> In the above example, the question can be answered, since Obama does have an oldest daughter. The correct answer option is 'Malia'. However, if you do not know the answer to this question, you should select 'Don't know'.

Cognitive Load Task. The use of cognitive resources has been advanced as a key feature of System 2 deliberation (Evans & Stanovich, 2013). Thus, to help prevent deliberation in the initial stage of our two-response paradigm, we imposed a concurrent cognitive load to participants during the trivia question answering. We used the dot memorization task (Miyake et al., 2001), which has been shown to successfully burden executive resources during verbal reasoning (e.g., De Neys & Schaeken, 2007; De Neys & Verschueren, 2006; Verschueren et al., 2004).

Before each trivia question, participants were presented with a 3 x 3 grid, in which four crosses were placed (Figure 1b). Participants were told that it was essential to memorize the location of the crosses while answering the questions. After their initial response, participants were shown four different matrices and they had to select the correct, to-be-memorized pattern. They then received feedback as to whether they chose the correct pattern. The load was only present during the initial response stage and not during the subsequent final response stage in which participants were allowed to deliberate (see Two-Response Paradigm section).

Procedure

One-Response Norming Study. To determine an appropriate deadline for the two-response paradigm we ran a traditional one-response version of the experiment without deadline or load (e.g., see Bago et al., 2021). The same material as in the main two-response experiment was used but participants only had to give a single answer for which they had all the time they wanted to deliberate, without any concurrent load. We recruited an independent sample of 50 online native English-speaking American (USA) participants (40 females, *M* age = 27 years, *SD* = 10 years) on the Prolific platform.

Results indicated that participants took on average 7.2 s (*SD* = 3.2 s) to provide correct responses to the anomaly problems (i.e., to respond “This question can’t be answered in this form”). In Experiment 1, we set the initial deadline at 5 s (see Two-response paradigm section), which corresponded to the first quartile of the correct, anomaly response latency of the one-response norming study. To test whether participants were under time pressure in the initial stage of the two-response paradigm, we contrasted latencies for anomaly correct responses in the one-response norming study and in the initial stage of the main two-response experiment, using a log-linear mixed effect model (see Table S12). Results indicated that participants responded significantly faster in the initial two-response stage (*M* = 3.8 s) than in the one-response norming study (*M* = 7.2 s), *t*(96.31) = -9.48, *p* < .001, Cohen’s *d* = 1.75, 95% CI [1.36, 2.14].

The one-response norming study also allowed us to check for a possible consistency confound in the two-response experiment. When people are asked to give two consecutive responses to the same question, they might want to appear consistent in their responses. This may prevent participants to correct their initial response, which would lead to an underestimation of the true correction rate and to a lower accuracy in the final stage of the two-response experiment. However, the results of a binomial generalized mixed-effects model showed that participants had virtually the same accuracy in the one-response norming study (*M* = 63.9%, *SD* = 17.6%) and in the final stage of the two-response paradigm (*M* = 64.9%, *SD* = 16.4%), *OR* = 1.06, *p* = .73, Cohen's *d* = 0.03, 95% CI [-0.13, 0.20] (see Table S13). This directly argues against a possible consistency confound in the two-response paradigm.

Two-Response Paradigm. We used a procedure similar to Bago and De Neys (2017). The experiment was run online on the Qualtrics platform. The task was introduced with the following instructions:

> **Please read these instructions carefully!**
>
> In this experiment you will have to respond to 20 multiple-choice trivia questions and a couple of practice questions.
>
> For every multiple choice question you will be presented with four answer options but **you can only pick one answer. Please respond as accurately as you can.**
>
> Some of the questions are impossible to answer. In that case, select the answer option: "This question can't be answered in this form."
>
> If you don't know the answer to a question, select the response option "Don't know".

Then, two examples were given to clarify the difference between the "Don't know" and the "This question can't be answered in this form" response options (see above). After this general introduction, participants were presented with a more specific instruction page about the procedure itself:

> Critically, in this study we want to know what your **initial, intuitive response** to the questions is and **how you respond after you have thought about these questions for some more time.**
>
> First, we want you to respond with the **very first answer that comes to mind**. You don’t need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible.
>
> To make sure that you answer as fast as possible, **a time limit was set for the first response**, which is going to be 5 seconds. When there is 1 second left, the background colour will turn to yellow to let you know that the deadline is approaching. Please make sure to **answer before the deadline passes.**
>
> Next, **the question will be presented again** and you can take all the time you want to actively reflect on it. Once you have made up your mind you give your **final response.**
>
> After you made your choice and clicked on it, you will be automatically taken to the next page.
>
> After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.

Participants then responded to two practice trivia questions to familiarize themselves with the deadline procedure. Next, they solved two practice load matrix problems without concurrent questions. Finally, at the end of the practice, they had to respond to the two earlier trivia questions using the complete two-response procedure (i.e., with deadline and load).

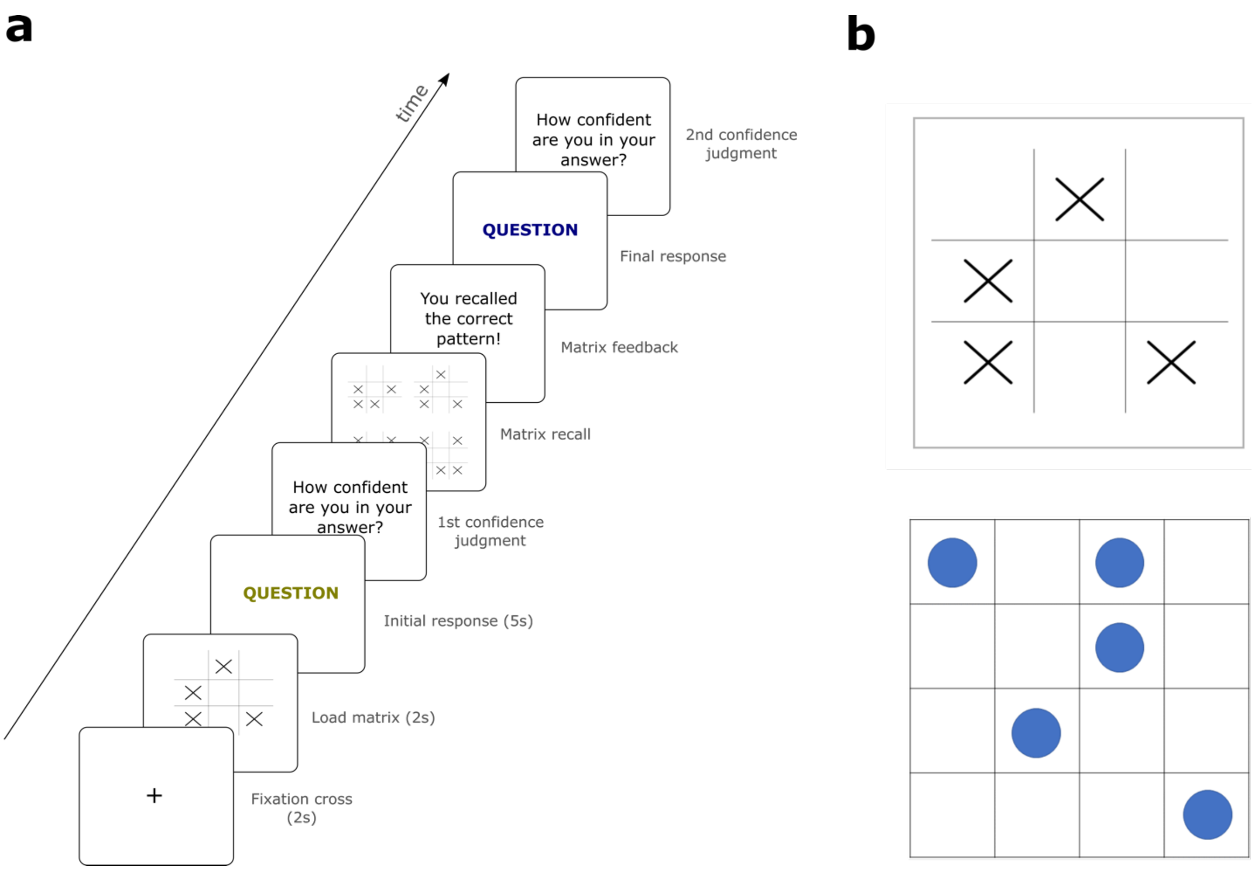

Figure 1. Experiment 1 trial sequence and examples of load patterns in Experiment 1-3. **a**) Example of one trial in Experiment 1. Participants had to respond to a trivia question twice, once with a deadline and a concurrent load and a second time without any constraint. **b**) Example of the to-be-memorized load patterns in Experiment 1-3 (upper panel) and Experiment 2 (lower panel).

Figure 1a shows a complete trial sequence for Experiment 1. At the beginning of each trial, a fixation cross was presented for 2 s. The load matrix was then presented for 2 s. After that, the question appeared. Participants had 5 s to respond. After 4 s, the background of the screen turned yellow to warn participants about the upcoming deadline. If they failed to provide an answer before the deadline, they were reminded to speed up and give an answer within the deadline on the next initial trials.

After the initial response, the question disappeared from the screen and participants had to enter their confidence in their response on a scale ranging from 0% to 100%, with the following instructions: “How confident are you in your answer? Please type a number from 0 (absolutely not confident) to 100 (absolutely confident)”. Then, participants had to select the to-be-memorized load pattern between four different load matrices. If they failed to select the correct pattern, they were reminded to make sure to remember the pattern correctly on the next trials.

The trivia question was then presented a second time, and participants could give their final response without any deadline, nor concurrent load. Once they had given their answer, they were automatically taken to the next page where they had to indicate their confidence level in their final answer.

To remind participants which question stage they were answering, the color of the answer options was green during the first response, and blue during the final response phase. In addition, we also added a reminder sentence under the question: “Please indicate your very first, intuitive answer.” and “Please give your final answer.”, respectively (see Supplementary Material B for the complete instructions).

After they had answered half of the questions, participants were allowed to take a short break. After they finished the experiment, they completed a page with standard demographic questions and were debriefed.

Exclusion Criteria. Participants failed to provide an initial response within the deadline on 6.8% of the trials and did not recall the correct load pattern on 12.7% of the trials. We removed any of the trials in which the deadline was missed or recall was inaccurate (or both) from our analyses because we cannot be sure that participants did not already deliberate to produce their initial response in these cases. Indeed, if participants did not respond within the deadline, they might have engaged in slow deliberation. Similarly, if they failed the load memorization task, we cannot guarantee that their cognitive resources were successfully burdened by the cognitive load.

Therefore, removing the trials that did not meet the inclusion criteria allowed us to be maximally sure that the initial responses were intuitive in nature.

Hence, a total of 18.2% of the trials were excluded, and we thus analyzed 1636 trials out of 2000. A slightly higher proportion of anomaly trials (20.8%) than no-anomaly trials (15.6%) were excluded from the analysis. On average, each participant contributed a total of 16.4 valid trials (out of 20, *SD* = 2.4 trials, range = 10–20).

In theory, these trial exclusions could have biased our results. More specifically, if the omitted trials were mainly incorrect initial responses, it may lead to an overestimation of correct intuitive responses. To confirm our results were robust to these exclusions, we reran all our main analyses while including these missed load and deadline trials. For missed deadline trials, where initial responses were absent, we conservatively coded these as incorrect responses for accuracy analyses. However, for our confidence analyses, we still had to exclude these missed deadline trials since we had no suitable values to replace them. In the missed load trials, both initial and final responses were recorded. The models are reported in Supplementary Material C. Including the previously excluded trials did not substantially alter our findings. We report the few instances where differences in significance emerged directly in the relevant results sections.

Statistical Analysis. The data were analyzed using (generalized) linear mixed models, which analyze data on a trial-by-trial basis while taking both participant and item variation into account. In contrast to traditional ANOVA analyses, mixed models enable more flexible analyses (as they do not require prior averaging), reduce Type I error, handle unbalanced datasets, and provide enhanced precision and generalization in prediction through partial pooling, where each individual participant's or item's data is informed by the complete dataset (Baayen et al., 2008).

When the random structure was unsupported or did not improve model fit, we used clustered standard errors to account for the non-independence of our data (Cameron & Miller, 2015).

To analyze dummy-coded dependent variables such as accuracy, we used binomial generalized mixed models (Bolker et al., 2009). For bounded dependent variables such as response confidence, we used mixed-effects beta regression (Verkuilen & Smithson, 2012). Given that beta regression can only handle data within the range of 0 to 1 (excluding these values), we adjusted the dependent variable to range between .005 to .995 and subsequently backtransformed it for interpretation on the original scale (Smithson & Verkuilen, 2006). For reaction times, we used linear mixed models along with a logarithmic transformation to accommodate non-normality. We subsequently backtransformed the reaction times to their original scale for interpretation. All estimated models can be found in Supplementary Material D with their corresponding random structure. Supplementary Material E specify how we found the best random structure for each analysis, how we evaluate significance in our different models and the contrast coding scheme we use depending on our analysis requirements.

The data were analyzed using the following R packages: *afex* (Singmann et al., 2015), *betareg* (Zeileis et al., 2016), *buildmer* (Voeten, 2020), *emmeans* (Lenth et al., 2019), *glmmTMB* (Brooks et al., 2017), *lme4* (Bates et al., 2014), *parameters* (Lüdecke et al., 2020), *sandwich* (Zeileis et al., 2020), and *tidyverse* (Wickham et al., 2019).

Results

Accuracy

Our first question of interest was whether people who provide a correct response to anomaly trivia questions after deliberation, can also provide a correct response to these questions when reasoning more intuitively in the initial response stage. Figure 2a provides a summary of the initial and final accuracy for the critical anomaly and control no-anomaly problems. For the control questions, participants performed very well both at the initial (*M* = 89.1%, *SD* = 14%) and the final (*M* = 92%, *SD* = 9.9%) response stages. These results indicate that overall participants knew the correct responses to the control questions and typically managed to generate them intuitively. The accuracy was lower for the anomaly questions, both in the initial response stage (*M* = 20.4%, *SD* = 22.6%) and in the final response stage (*M* = 35.9%, *SD* = 30.6%). The results for the anomaly questions thus indicate that our items succeeded in eliciting the Moses illusion. More importantly, they also show that participants managed to give a significant number of correct answers at the initial response stage for anomaly questions, although they performed better at the final response stage.

Table 2

Binomial generalized mixed-effects model on accuracy as a function of anomaly presence and response stage in Experiment 1, using sum coding

accuracy ~ 1 + anomaly + response stage + anomaly:response stage + (1 + anomaly | subject) + (1 + anomaly | item)

Fixed effects

Predictors	df	χ^2	p.value	Cohen's d
Anomaly	1	41.68	**< .001**	-1.39
Response stage	1	51.98	**< .001**	0.24
Anomaly:Response stage	1	10.74	**.002**	0.11

Random effects

Group	Parameter	SD
Subject	(Intercept)	1.30
Item	(Intercept)	0.84
Subject	Anomaly	0.86
Item	Anomaly	0.87
Subject	Cor (Intercept x Anomaly)	0.35
Item	Cor (Intercept x Anomaly)	-0.50

Model fit

Metric	R^2 (Marginal)	R^2 (Conditional)
	0.48	0.76

N	Subject	Item	Observations
	100	20	3,272

Note. p-values were obtained through bootstrapping on the comparison between a full model and a reduced model without the variable of interest, using type III sum of squares.

To test these results statistically, we built a binomial generalized mixed-effects model with accuracy as the dependent variable (see Table 2). Anomaly presence (control no-anomaly; anomaly), response stage (initial; final), and their interaction were included as fixed effects using sum coding. There was a significant main effect of anomaly on accuracy, as well as response stage. Finally, the difference between initial and final accuracy was higher for anomaly questions compared to control no-anomaly questions, as indicated by the response stage by anomaly interaction.

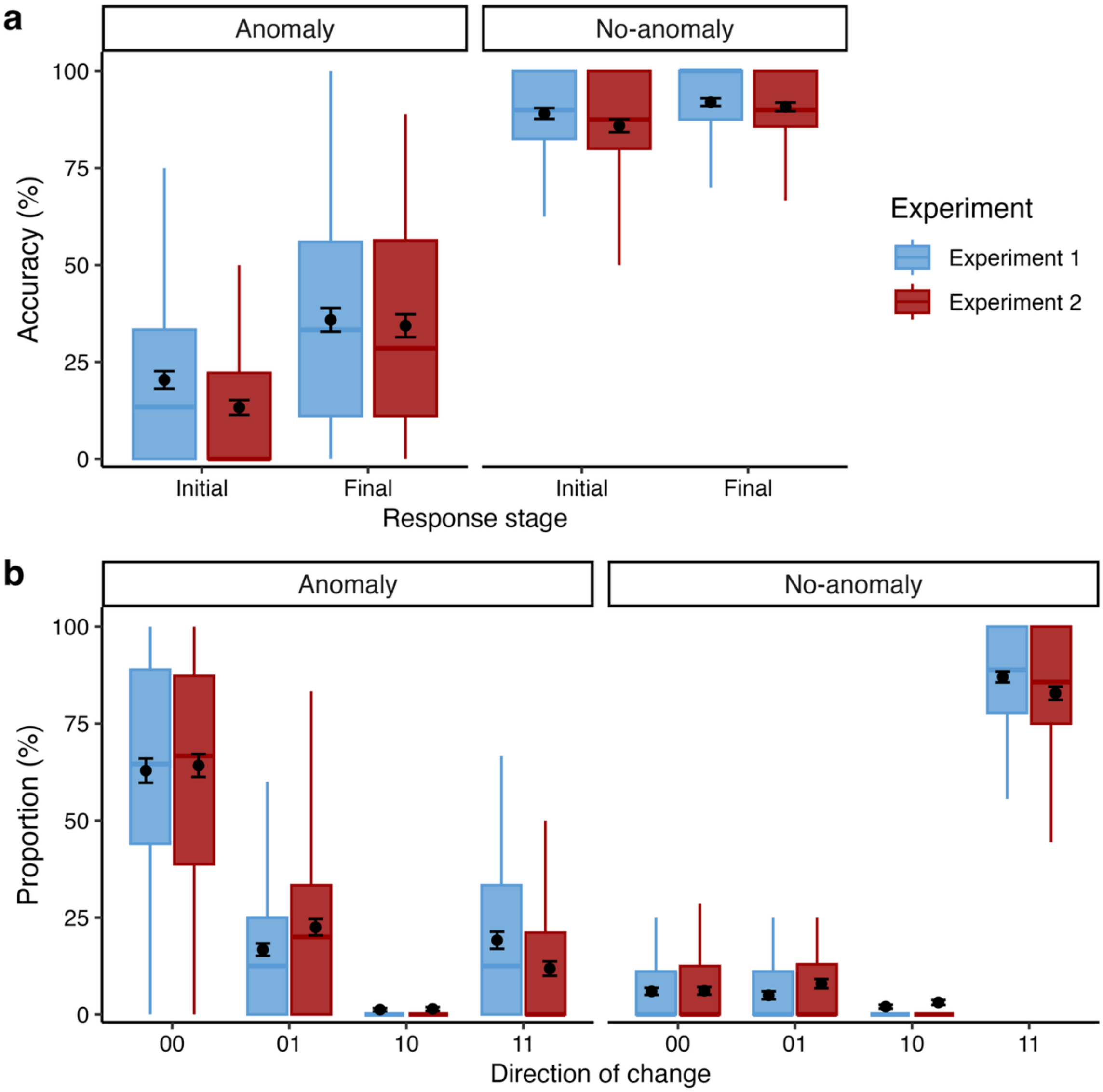

Figure 2. Accuracy and Direction of Change in Experiment 1 and Experiment 2. **a**) Response accuracy at anomaly and control no-anomaly trials as a function of response stage. **b**) Proportion of each direction of change category at anomaly and control no-anomaly trials; “00” = incorrect initial and incorrect final response; “01” = incorrect initial and correct final response; “10” = correct initial and incorrect final response; “11” = correct initial and correct final response. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile range. Overlaid black dots represent the mean and black error bars are standard errors of the mean.

Due to our multiple-choice response format and the high demands of the initial response stage, one might argue that participants were simply guessing and responding randomly at that stage. However, the remarkably high accuracy on control no-anomaly problems (M = 89.1%) challenges this potential confound, suggesting that participants did overall not blindly guess during the initial response stage. Another argument against guessing stems from the type of errors observed in the anomaly questions. If participants were responding randomly in the intuitive block, their mistakes should be evenly distributed among the undistorted answer option (e.g., "Two" in the original Moses illusion), the filler option (e.g., "Three" in our filler response option), and the "Don't know" response option. However, our findings reveal that among the initial incorrect responses to anomaly problems (i.e., 79.6% of the total number of responses for the initial anomaly problems), the overwhelming majority of errors occurred because participants selected the undistorted answer option (86%[2]) rather than the filler option (5.3%) or the "Don't know" response option (8.7%).

Direction of Change

To better understand how people changed (or did not change) their answers after deliberation, we performed a direction of change analysis by looking into how accuracy changed from the initial to the final response stage on every trial (Bago & De Neys, 2017). For each trial, participants have either an accuracy of "1" (i.e., correct response) or "0" (i.e., incorrect response) in each of the two response stages (i.e., initial and final). Hence, there are four possible directions

[2] This figure is very close to the proportion of undistorted answer options for incorrect anomaly answers in the final response stage (85.2%). However, to prevent confusion, it is important to note that this proportion is conditioned on incorrect responses, which were more frequent in the initial response stage than in the final response stage. This accounts for the apparent lack of difference in the undistorted response proportions across the two stages. In absolute terms, the proportion of undistorted responses decreased from 78.7% in the initial response stage to 74.1% in the final response stage in Experiment 1, in line with our accuracy results. Similarly, in our other experiments, the proportion of undistorted responses decreased from 78.7% to 74.4% in Experiment 2, and from 72% to 67.8% in Experiment 3.

of change: “00” (incorrect initial and incorrect final response), “01” (incorrect initial and correct final response), “10” (correct initial and incorrect final response) and “11” (correct initial and correct final response).

Figure 2b shows the mean direction of change frequencies as a function of anomaly presence. As expected, the vast majority of control no-anomaly questions yielded “11” responses (87%), followed by “00” (6%), “01” (5%) and “10” responses (2%). Regarding the critical anomaly questions, the majority of “00” responses (62.9%) is consistent with the literature, as it shows that participants tend to fall for the illusion even when they are allowed to deliberate. There were slightly more “11” responses[3] (19.2%) than “01” responses (16.7%), whereas the “10” response pattern was rare (1.2%).

Testing the corrective deliberation assumption of dual process accounts of the Moses illusion requires to concentrate on trials where participants gave a correct answer in the final response stage. To get a more direct measure of how the proportion of “11” responses compared to that of “01” responses, we computed the mean “non-correction rate” across participants for the anomaly questions (i.e., proportion 11/11+01; Bago & De Neys, 2017). Note that here we computed the non-correction rate for each participant before computing the mean of these individual non-correction rates. We thus excluded the participants who never gave a final correct answer to anomaly questions (i.e., no “11” or “01” response whatsoever; n = 21). This measure indicates the proportion of correct final answers that were already correct in the initial response stage. Put differently, it shows the proportion of trials for which participants did not need to deliberate to find

[3] As one reviewer noted, given that with our 4-response option multiple-choice format there is a 6.25% chance to observe a “11” pattern under pure random responding, one can also contrast the observed “11” frequency against this benchmark to test for a guessing confound. One-sample t-tests (one-tailed) showed that the “11” frequencies differed significantly from this chance level in Experiment 1 and in Experiment 2 (Experiment 1: $p < .001$; Experiment 2: $p = .002$).

the correct answer. If deliberate correction is critical for correct responding, it should be at 0%. Instead, the mean non-correction rate for anomaly questions was 46.1% (*SD* = 34.1%). It means that on average, when participants managed to give a correct answer to an anomaly question at the final stage, they already gave a correct answer at the initial stage about half the time. The full distribution of the individual non-correction rates is reported in Supplementary Material F.

To obtain a statistical estimate of the non-correction rate, we selected the “01” and “11” anomaly responses. We then created a dummy variable, coding it as 0 for “01” responses and 1 for “11” responses. Subsequently, we built a binomial mixed-effects model, using this dummy variable as our dependent variable (see Table S16). This allowed us to directly derive the non-correction rate using the estimated mean from the model. The resulting estimate was very similar to the descriptive mean, yielding an estimated non-correction rate of 49.1%, 95% CI [36.8, 60.2]. In summary, although deliberative correction did occur in Experiment 1, in many cases participants were able to generate the correct response when deliberation was minimized in the initial response stage.

Confidence Ratings

Error Sensitivity. Our second question of interest was to see whether people who give incorrect responses to anomalous trivia questions in the intuitive response stage show some sensitivity to the presence of the anomaly and detect that their answer is questionable. Participants who fall prey to the illusion typically answer with the undistorted response (i.e., “Two” in the Moses Illusion). Whereas this answer is correct for the undistorted control problems, it is obviously incorrect for the distorted anomaly problems. Hence, by contrasting participants’ response confidence for correctly solved control no-anomaly problems and incorrectly solved anomaly problems we can test whether they display some basic error sensitivity on their initial answers. If incorrect responders do not register the anomaly, they should not process the two problem versions any differently and should be equally confident about their answers. If incorrect responders show increased doubt when they err, this indicates that they detect that their response is questionable and–despite their incorrect answer–show some minimal sensitivity to the presence of the anomaly.

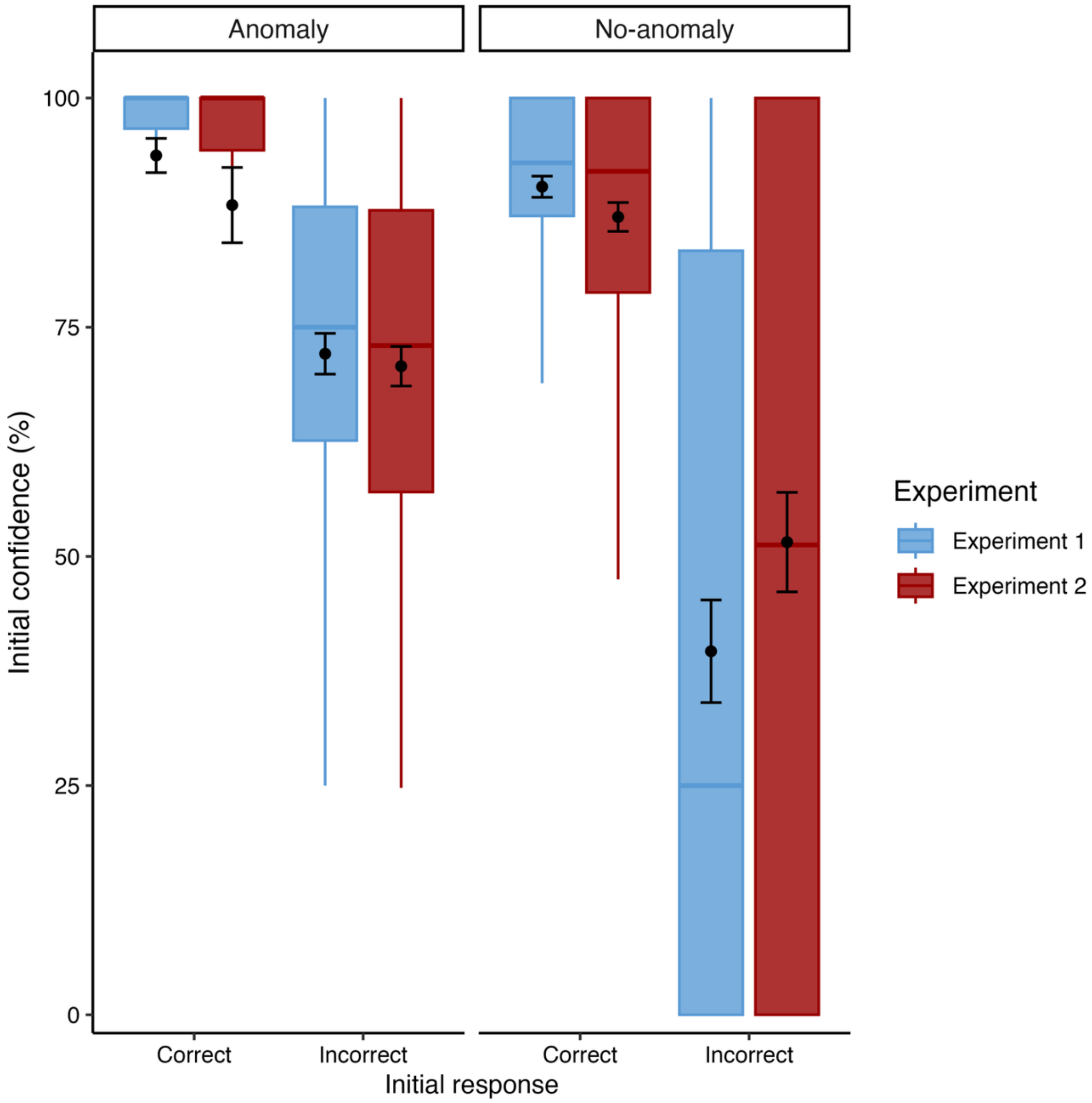

Figure 3. Confidence ratings at anomaly and control no-anomaly trials in the initial response stage as a function of accuracy in Experiment 1 and Experiment 2. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile range. Overlaid black dots represent the mean and error bars are standard errors of the mean.

Figure 3 displays the mean initial confidence ratings for anomaly and control no-anomaly trials as a function of accuracy (i.e., correct vs. incorrect responses). We built a beta generalized mixed-effects model on the initial confidence as a function of anomaly and accuracy using dummy coding (see Table 3). The results indicated that the confidence ratings for anomaly incorrect responses (*M* = 72.1%) were significantly lower than the confidence ratings for control correct responses (*M* = 90.3%). Participants thus showed increased response doubt when they were making a mistake which suggests they were detecting that their answer was questionable, even in the initial response stage when deliberate reflection was minimized. The full distribution of the individual initial error sensitivity measures is reported in Supplementary Material F. One may argue that these findings may be better explained by the fact that we coded the (very rare) “Don’t know” responses as incorrect (see Method section). As these responses may receive lower confidence ratings on average, they could be responsible for our confidence findings. To rule out this alternative explanation, we also performed all our confidence analyses without the “Don’t know” responses. Overall, the confidence results were strongly consistent whether or not these responses were included; any discrepancies are reported in the relevant sections.

Table 3

Mixed-effects beta-regression on initial confidence ratings as a function of anomaly and accuracy in Experiment 1, using dummy coding

confidence ~ 1 + anomaly + accuracy + anomaly:accuracy + (1 | item)

Fixed effects

Predictors	OR	95% CI	p.value	Cohen's d
Intercept (no-anomaly correct)	3.43	[3.01, 3.88]	**< .001**	0.68
Anomaly incorrect	0.66	[0.57, 0.75]	**< .001**	-0.23
Anomaly correct	1.18	[0.96, 1.42]	.12	0.09
No-anomaly incorrect	0.20	[0.15, 0.26]	**< .001**	-0.88

Random effects

Group	Parameter	SD
Item	(Intercept)	0.16

Model fit

Metric	R² (Marginal)	R² (Conditional)
	0.44	0.51

N	Subject	Item	Observations
	100	20	1,636

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

For completeness, as shown in Figure 3 and confirmed in the model, correct responses to anomaly problems were not associated with a confidence decrease relative to correct responses on control problems. In fact, participants were slightly more confident when they successfully avoided the illusion (*M* = 93.7%) than when responding correctly to control problems (*M* = 90.3%), though this difference was not statistically significant. Finally, confidence was significantly lower in the incorrect control no-anomaly problems compared to correct response on control problems (*M* = 39.7%).

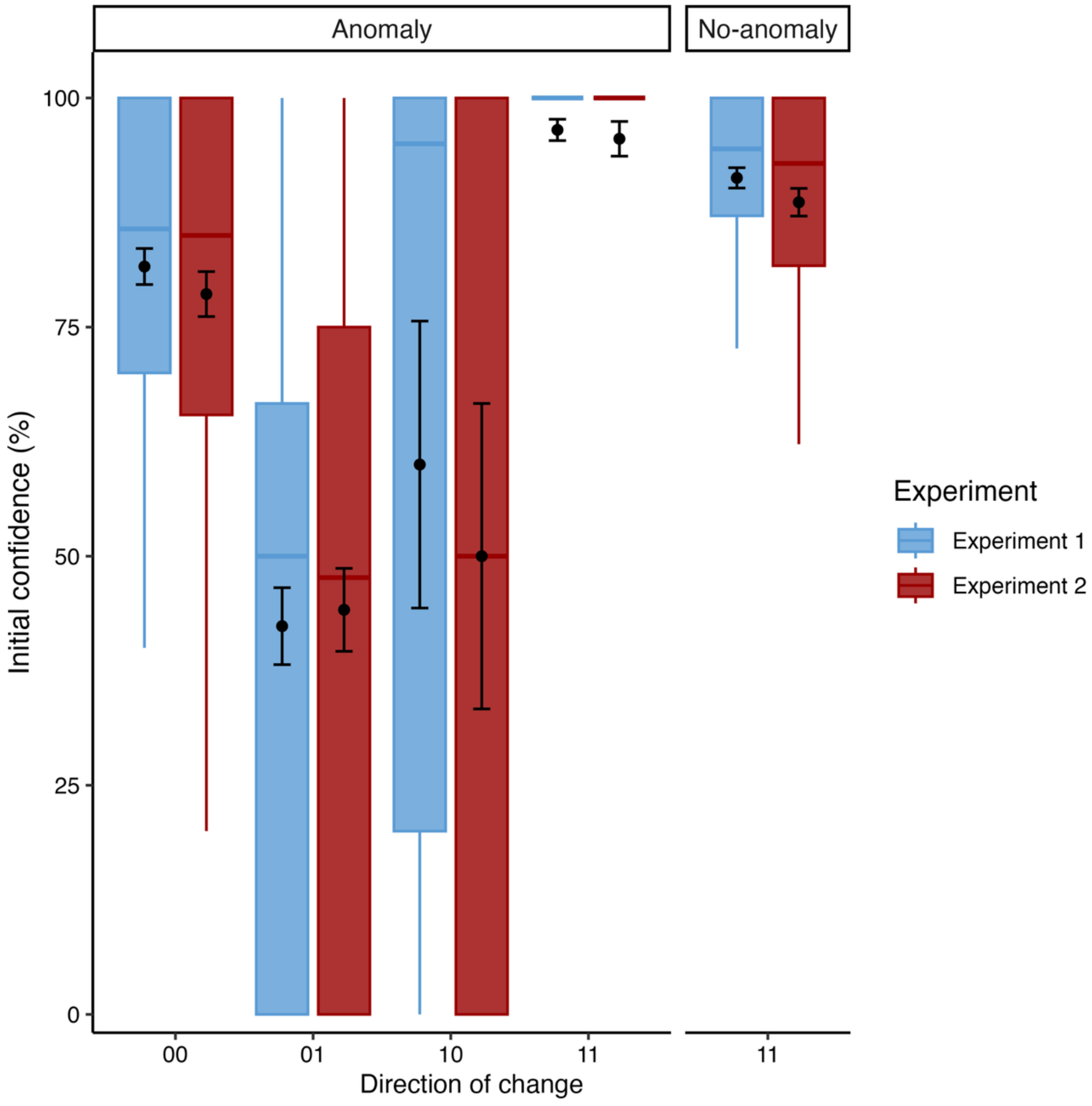

Figure 4. Initial confidence ratings at anomaly trials as a function of each direction of change category in Experiment 1 and Experiment 2; the right panel displays the "11" control no-anomaly responses, which was coded as the intercept (reference category) in the model; "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct initial and correct final response. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile range. Overlaid black dots represent the mean and error bars are standard errors of the mean.

Direction of Change. For exploratory purposes, we also analyzed the confidence ratings as a function of direction of change. A classic finding in the reasoning field is that trials where participants change their initial response (i.e., “01” or “10” response categories) tend to show lower initial response confidence than trials where participants stick to their initial answer (i.e., “11” or “00” responses, e.g., Bago & De Neys, 2017; Thompson et al., 2011). This low confidence (or “Feeling of Rightness”) is considered as a key determinant of deliberate answer change (Bago & De Neys, 2017; Thompson et al., 2011). Figure 4 shows the mean initial confidence ratings for each direction of change category for anomaly problems. We clearly replicate the reasoning pattern and find lower initial confidence for the change (“01” and “10”) than no change (“11” and “00”) categories.

Table 4

Beta-regression results contrasting the initial confidence ratings for "11" control no-anomaly trials with anomaly trials for each direction of change category in Experiment 1, using dummy coding

confidence ~ 1 + response category + (1 | item)

Fixed effects

Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly "11")	3.94	[3.50, 4.42]	**< .001**	0.76
Anomaly "11"	1.18	[0.96, 1.44]	.12	0.09
Anomaly "00"	0.79	[0.70, 0.90]	**< .001**	-0.13
Anomaly "01"	0.21	[0.17, 0.27]	**< .001**	-0.85
Anomaly "10"	0.45	[0.18, 0.95]	**.04**	-0.45

Random effects

Group	**Parameter**	**SD**
Item	(Intercept)	0.16

Model fit

Metric	**R^2 (Marginal)**	**R^2 (Conditional)**
	0.54	0.61

N	**Subject**	**Item**	**Observations**
	100	20	1,525

Note. p-values and confidence intervals were obtained using parametric bootstrap. OR = odds ratio. An OR superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence relative to the to the "11" no-anomaly control condition.

To analyze the results statistically, we used a mixed-effect beta-regression model (see Table 4). We built a model contrasting the initial confidence rating for "11" control no-anomaly trials (which served as our baseline) to the initial confidence rating of each direction of change for anomaly trials using dummy coding (see Bago & De Neys, 2017 for a similar analysis).

The direction-of-change categories associated with the largest significant decreases in confidence compared to the baseline were the "01" category and the very rare "10" category[4]. A smaller but still significant decrease was observed for the "00" category[5]. In contrast, the "11" anomaly category showed a non-significant increase in confidence relative to the control trials. To further explore the role of response change on initial confidence, we compared trials in which participants changed their response between the initial and final stages ("01" and "10" categories) to those in which they did not change their response ("00" and "11" categories). Replicating previous findings, a post-hoc contrast showed that initial confidence was significantly lower in change trials than in no-change trials, *OR* = 3.15, 95% CI [2.10, 5.15], $p < .001$, Cohen's $d = 0.63$.

Illusion Strength Analysis. Overall, our anomaly items managed to trigger semantic illusions and we observed low average accuracy across our problems. However, not all individual items triggered the illusion to the same extent. For some problems the intuitive pull of the impostor word seemed stronger than others. This naturally occurring variance in "illusion strength" (i.e., how difficult it is to spot the illusion as operationalized by response accuracy) can be used to further validate our confidence findings. Our overall evidence for intuitive error detection suggests that when responding incorrectly to the question "In the biblical story, how many animals of each

[4] However, note that the difference between the baseline and the anomaly "10" category was no longer significant ($p = .08$) when including the trials with a missed cognitive load (see Supplementary Material C for the full model).

[5] The difference between the baseline and the anomaly "00" category no longer reached significance ($p = .07$) when the "Don't know" responses were excluded from the analysis.

kind did Moses take on the Ark?", the correct concept of "Noah" is also activated on some level. However, the strength of this correct intuition may differ across items. The stronger the illusion (i.e., the lower the correct intuition's strength / the lower the item's accuracy), the harder it will be to spot the anomaly, and the less likely that people will show error sensitivity and doubt their answer. In the reasoning field, such a link between illusion strength and error detection has already been established (e.g., Bago & De Neys, 2020; De Neys & Pennycook, 2019; Pennycook et al., 2015).

Interestingly, for correctly solved anomaly problems one can make the exact opposite prediction with respect to response confidence. That is, our overall analysis indicated that on average correct responders did not doubt their answer. However, if an illusion is particularly strong, even correct responders may feel more conflicted and less confident about their answer. This pattern has also been observed in the reasoning field (e.g., Bago & De Neys, 2020; De Neys & Pennycook, 2019; Pennycook et al., 2015).

Here, we ran a post-hoc analysis to explore these hypotheses. As an exploratory first test of these two predictions, we computed an "illusion strength" measure for each of our 20 items. For each question, we calculated the difference between the mean accuracy of the control no-anomaly version and the mean accuracy of the anomaly version in the initial, intuitive response stage. The mean accuracy of the control version thus served as a baseline: The lower the average anomaly version accuracy in comparison with the control version, the stronger the illusion. An items' illusion strength thus reflects its capacity to elicit the illusion in the initial stage at the group level. To recap, we expect that as illusion strength increases (i.e., the correct intuition decreases), participants will show more confidence in their errors (i.e., less error detection) and less confidence in correct responses.

Table 5

Beta-regression predicting initial confidence ratings in Experiment 1 as a function of standardized illusion strength, response group, and their interaction, using dummy coding

confidence ~ 1 + group + illusion strength + illusion strength:group

Predictors	OR	95% CI	Z value	p.value	Cohen's d
Intercept (no-anomaly correct)	3.59	[3.49, 3.69]	24.8	**<.001**	0.71
Anomaly correct	1.10	[0.99, 1.21]	1.71	.09	0.05
Anomaly incorrect	0.63	[0.51, 0.74]	-8.12	**<.001**	-0.26
Illusion strength	1.15	[1.11, 1.18]	7.65	**<.001**	0.08
Anomaly correct:Illusion strength	0.83	[0.72, 0.93]	-3.49	**<.001**	-0.11
Anomaly incorrect:Illusion strength	1.15	[1.05, 1.26]	2.64	**.008**	0.08

Model fit

Metric	Pseudo R²
	0.11

N	Subject	Item	Observations
	100	20	1,542

Note. p-values were obtained using Wald Z-tests with clustered standard errors by participants. OR = odds ratio. An OR superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

To analyze the data, we used a beta-regression on the initial confidence with clustered standard errors by participant using dummy coding (see Table 5). As fixed factors, we entered a variable which we will refer to as "response group", illusion strength (standardized), and their interaction. The "response group" variable coded whether a given data point was a control trial on which the correct response was selected (intercept of the model), an anomaly trial on which the correct response was selected, or an anomaly trial on which the incorrect response was selected. Figure 5 plots the result of the regression. Note that a regression line parallel to the correct control baseline (dashed line) would indicate that confidence is not modulated by illusion strength.

Critically, the interaction term between illusion strength and the response groups was significant both for correct anomaly answers and for incorrect anomaly answers. As Figure 5 indicates, as illusion strength increased (i.e., the alleged correct intuition decreased), confidence decreased for correct anomaly responses and increased for incorrect anomaly responses. Hence, as could be expected, error sensitivity became less pronounced for "harder" problems whereas correct responses to these problems were doubted more.

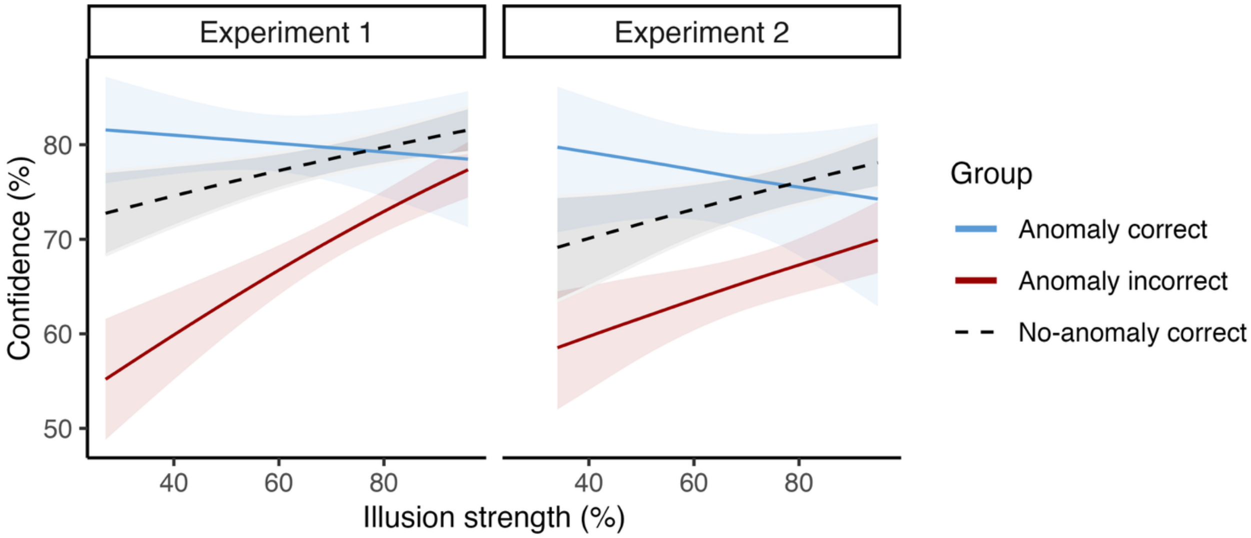

Figure 5. Regression results of initial confidence as a function of illusion strength for control no-anomaly correct (baseline), anomaly correct and anomaly incorrect responses. Illusion strength = mean initial no-anomaly accuracy − mean initial anomaly accuracy for each item. The shaded bands are 95% confidence bands.

Experiment 2

The results of our first experiment challenge the dual process account of semantic illusions. When participants managed to give a correct answer to an anomaly problem at the final stage, they already gave a correct answer at the initial stage about half the time (mean non-correction rate = 46.1%). This result questions the corrective deliberation assumption in dual process accounts of semantic illusions. In addition, when participants failed to answer correctly in the initial response stage, they were still able to detect their error to some extent, as indicated by a decrease in their confidence.

In Experiment 2, we introduced methodological refinements to test the robustness of the findings. Although in Experiment 1 we combined three validated procedures (instructions, time pressure, and concurrent load) to minimize deliberation in the initial response stage, one concern is that the procedure was not stringent enough. If participants already deliberated in the initial response stage, this could explain the high non-correction rate and initial error sensitivity. To rule out this concern, we used more extreme load and time pressure manipulations (for a similar approach, see Bago & De Neys, 2019a).

Method

Transparency and openness.

The research question and experimental design were preregistered on the OSF platform (https://osf.io/295bf). No specific analyses were preregistered.

Participants

We recruited 100 online participants (77 females, *M* age = 35.1 years, *SD* = 15.4 years) on the Prolific platform. Only native English-speaking American (USA) participants were allowed to participate in the experiment. Participants were paid £1.2 for their 12 minutes of participation. Among them, 44 reported high school as their highest educational level, 1 less than high school, and 55 a higher education degree.

Materials and Procedure

Except for the initial response deadline and the load task, the materials and the procedure were the same as in Experiment 1.

Response Deadline. In Experiment 1, the initial response deadline was set to 5 seconds, based on our one-response norming study. The results of Experiment 1 indicated that on average participants respected the instructions and overall responded before the deadline. To further minimize the possibility that participants engage in deliberation during the initial response stage, we decided to use a more stringent time limit. The average correct, initial anomaly response latency in Experiment 1 was 3.8 s. Based on this result, we decided to round this value to the nearest integer (to give participants some minimal leeway) and decreased the deadline further to 4 s in Experiment 2. The screen turned yellow 1 s before the deadline to urge participants to enter their response.

To check whether the time pressure had increased between Experiment 1 and Experiment 2, we contrasted the response latencies in the initial response stage of the two experiments (see Table S14). A log-linear mixed effects model showed that participants responded significantly faster in Experiment 2 (*M* = 2.9 s) than in Experiment 1 (*M* = 3.2 s), *t*(197.94) = -4.19, *p* < .001, Cohen's *d*

= 0.43, 95% CI [0.23, 0.64]. However, note that although we decreased the deadline by one entire second between the two experiments, the mean initial latency difference was only 0.3 s. This indicates that participants were already responding near the minimal threshold in Experiment 1.

Load Task. In Experiment 2, we also increased the cognitive load during the initial response stage. In Experiment 1, participants had to memorize a complex four-cross pattern in a 3 x 3 grid. In Experiment 2, we presented a five-dot pattern in a 4 x 4 grid (Figure 1b); e.g., Bialek & De Neys, 2017; Trémolière & Bonnefon, 2014). This more extreme load has been shown to further burden participants' cognitive resources compared to the load task we used in Experiment 1 (Trémolière et al., 2012). Except for the more demanding five-dot patterns, the load task was the same as in Experiment 1.

Exclusion Criteria. Participants failed to provide an initial response within the deadline in 14.8% of the trials and did not recall the correct dot pattern in 20.3% of the trials. A total of 31.6% of the trials were excluded, and we thus analyzed 1369 trials out of 2000. A slightly higher proportion of anomaly trials (33.7%) than no-anomaly trials (29.4%) were excluded from the analysis. On average, each participant contributed a total of 13.7 valid trials (out of 20, *SD* = 3.3 trials, range = 3–19).

Note that this higher proportion of excluded trials in Experiment 2 (i.e., 31.6% vs. 18.2% in Experiment 1) was to be expected, as we used a very stringent deadline and a more demanding load to be maximally sure that participants could not engage in slow deliberation during the initial response stage. However, since we only discarded individual trials (rather than participants), this higher exclusion rate should not give rise to confounding selection effects at the participant level (e.g., Bouwmeester et al., 2017). Furthermore, significance did not change when including the trials where participants missed the deadline or failed the cognitive load task (see Supplementary Material C).

Results and Discussion

Accuracy

Figure 2a summarizes the initial and final accuracies as a function of anomaly presence. These parallel the Experiment 1 findings. For the control no-anomaly questions, participants performed very well both at the initial (*M* = 85.9%, *SD* = 16.6%) and the final (*M* = 90.8%, *SD* = 11.2%) response stages. The accuracy was lower for the anomaly questions, again both in the initial (*M* = 13.3%, *SD* = 19%) and final response stage (*M* = 34.4%, *SD* = 29.3%).

Table 6

Binomial generalized mixed-effects model on accuracy as a function of anomaly presence and response stage in Experiment 2, using sum coding

accuracy ~ 1 + anomaly + response stage + anomaly:response stage +
(1 + anomaly | subject) + (1 + anomaly || item)

Fixed effects

Predictors	**df**	χ^2	**p.value**	**Cohen's d**
Anomaly	1	44.08	< **.001**	-1.27
Response stage	1	87.23	< **.001**	0.32
Anomaly:Response stage	1	22.31	< **.001**	0.17

Random effects

Group	**Parameter**	**SD**
Subject	(Intercept)	1.04
Item	(Intercept)	0.71
Subject	Anomaly	0.81
Item	Anomaly	0.75
Subject	Cor (Intercept x Anomaly)	0.55

Model fit

Metric	**R² (Marginal)**	**R² (Conditional)**
	0.54	0.69

N	**Subject**	**Item**	**Observations**
	100	20	2,738

Note. p-values were obtained through bootstrapping on the comparison between a full model and a reduced model without the variable of interest, using type III sum of squares.

A binomial generalized mixed-effects model on accuracy including anomaly presence (control no-anomaly; anomaly), response stage (initial; final) and their interaction as fixed effects using sum coding revealed a significant main effect of anomaly on accuracy, as well as response stage (see Table 6). The difference between initial and final accuracy was higher for anomaly questions compared to control questions, as indicated by the response stage by anomaly interaction.

Given the increased demands of the deadline and cognitive load in Experiment 2, one may contend that participants were simply overall guessing in the initial response stage. Once again, the very high accuracy on control no-anomaly problems (*M* = 85.9%) argues against this potential guessing confound. Additionally, the nature of errors made in anomaly problems during the initial response stage offers further evidence against participants responding at random. If participants were solely guessing at this stage, their mistakes would be evenly distributed among the undistorted answer option (e.g., "Two" in the original Moses illusion), the filler option (e.g., "Three" in our filler response option), and the "Don't know" response option. However, our findings reveal that among the initial errors made in response to anomaly problems (representing 86.7% of the total number of responses for the initial anomaly problems), the majority occurred because participants selected the undistorted answer option (82.2%), rather than the filler (8.9%) or the “Don’t know” response options (8.9%).

Overall, the pattern of results was similar in Experiment 1 and 2. Participants’ performance on anomaly problems is better at the final response stage, but they still manage to give correct answers at the initial response stage.

Direction of Change

To better understand how people changed (or did not change) their answers after deliberation, we once again performed a direction of change analysis by looking into how accuracy changed from the initial to the final response stage on every trial. As a reminder, there are four possible directions of change: “00” (incorrect initial and final response), “01” (incorrect initial and correct final response), “10” (correct initial and incorrect final response), and “11” (correct initial and final response). Figure 2b gives an overview of the results. As in Experiment 1, the majority of control no-anomaly questions yielded “11” responses (82.8%), followed by “01” (8%), “00” (6.1%) and “10” responses (3.1%). Similarly, there was a majority of “00” responses (64.2%) for anomaly questions. However, “01” responses were more frequent (22.5%) than “11” responses (11.9%) in Experiment 2. The “10” response pattern was again rare (1.4%).

In order to get a more direct measure of how the proportion of “11” responses compared to that of “01” responses, we again computed the mean “non-correction rate” across participants who had managed to give at least one correct answer to an anomaly question at the initial or final stage (n = 75). The non-correction rate measures how the proportion of “11” responses compares to that of “01” responses (i.e., proportion 11/11+01). The mean non-correction rate for anomaly questions was 28.1% (*SD* = 34.1%). Put differently, when participants managed to give a correct answer to an anomaly question at the final stage, they already gave a correct answer at the initial stage 28.1% of the time. To statistically estimate the non-correction rate in Experiment 2 and compare it with that of Experiment 1, we once again created a dummy variable coded as 0 for “01” responses and 1 for “11” responses. Subsequently, we constructed a binomial mixed-effects model, using this dummy variable as our dependent variable and experiment (Experiment 1; Experiment 2) as a fixed effect (see Table S16). The estimated non-correction rate derived from the model was 28.7%, 95% CI [18.8, 39.8], which proved to be significantly lower than the 46.1% non-correction rate observed in Experiment 1, $p < .001$, Cohen's $d = -0.48$, 95% CI [-0.74, -0.22]. In summary, our data indicates that it was still possible to generate the correct answer intuitively in Experiment 2. However, the correction of an erroneous intuitive answer was more frequent than a correct intuitive answer.

Confidence Ratings

Error Sensitivity. To test whether incorrect anomaly problem responders detected their error, we again contrasted the confidence ratings of the correctly solved control no-anomaly problems and the incorrectly solved anomaly problems in the initial response stage of the paradigm. As Figure 3 shows, we replicated the findings of Experiment 1. A beta generalized mixed-effects model on the initial confidence as a function of anomaly and accuracy using dummy coding (see Table 7) showed that the initial confidence ratings on incorrectly solved anomaly problems (M = 70.7%) were significantly lower than those for correctly solved control problems (M = 87%). Participants were thus able to detect that their answer was questionable, even in the initial response stage when deliberate reflection was minimized.

Table 7

Mixed-effects beta-regression on initial confidence ratings as a function of anomaly and accuracy in Experiment 2, using dummy coding

confidence ~ 1 + anomaly + accuracy + anomaly:accuracy + (1 | subject)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly correct)	2.95	[2.64, 3.31]	**< .001**	0.60
Anomaly incorrect	0.65	[0.56, 0.75]	**< .001**	-0.24
Anomaly correct	1.08	[0.83, 1.42]	.58	0.04
No-anomaly incorrect	0.33	[0.25, 0.43]	**< .001**	-0.61

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.26

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.27	0.44

N	**Subject**	**Item**	**Observations**
	100	20	1,369

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

As in Experiment 1, for correct anomaly responses we did not observe a similar confidence decrease when contrasting the control no-anomaly and the anomaly problems (see Table 7). When participants avoided the illusion and responded correctly, they were highly confident that their answer was indeed correct (*M* = 88.3% for correct anomaly problems vs. 87% for correct control problems). Finally, confidence was low in the incorrect no-anomaly problems (*M* = 51.6%).

Direction of Change. For exploratory purposes, we also analyzed the confidence ratings as a function of direction of change in Experiment 2. As showed in Figure 4, we replicate the results of Experiment 1, by finding lower initial confidence for the change ("01" and "10") than no change ("11" and "00") categories.

Table 8

Beta-regression results contrasting the initial confidence ratings for "11" control no-anomaly trials with anomaly trials for each direction of change category in Experiment 2, using dummy coding

confidence ~ 1 + response category + (1 | subject)

Fixed effects

Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly "11")	3.39	[3.03, 3.82]	**< .001**	0.67
Anomaly "11"	1.15	[0.86, 1.50]	.35	0.08
Anomaly "00"	0.79	[0.68, 0.91]	**.004**	-0.13
Anomaly "01"	0.25	[0.20, 0.32]	**< .001**	-0.77
Anomaly "10"	0.29	[0.13, 0.67]	**.002**	-0.68

Random effects

Group	**Parameter**	**SD**
Subject	(Intercept)	0.21

Model fit

Metric	**R^2 (Marginal)**	**R^2 (Conditional)**
	0.47	0.58

N	**Subject**	**Item**	**Observations**
	100	20	1,247

Note. p-values and confidence intervals were obtained using parametric bootstrap. OR = odds ratio. An OR superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence relative to the to the "11" no-anomaly control condition.

To analyze the results statistically, we used a mixed-effect beta-regression model (see Table 8) contrasting the initial confidence rating for “11” control no-anomaly trials (which served as our baseline) to the initial confidence rating of each direction of change for anomaly trials using dummy coding. As in Experiment 1, the direction of change categories associated with the biggest confidence decrease compared to the control trials were the “01”, and the –very rare– “10” categories, while the decrease was smaller for the “00” category. Finally, the “11” category once again showed a non-significant increase in confidence relative to the control trials. As in Experiment 1, a post-hoc contrast confirmed that initial confidence was significantly lower in change (i.e., “01” and “10” categories) trials than in no-change (i.e., “11” and “00” categories) trials, *OR* = 3.50, 95% CI [2.30, 5.54], *p* < .001, Cohen’s *d* = 0.69.

Illusion Strength Analysis. As in Experiment 1, we also performed an illusion strength analysis. For each item, we computed the illusion strength (i.e., how difficult it is to spot the illusion) by subtracting the mean initial accuracy of the anomaly version from the mean initial accuracy of the control no-anomaly version. The higher the difference between the two, the stronger we assume the illusion to be. As in Experiment 1, we predict that as illusion strength increases, the experienced confidence will decrease for correct responses and increase for incorrect responses (i.e., less error detection).

Table 9

Mixed-effects beta-regression predicting initial confidence ratings in Experiment 2 as a function of standardized illusion strength, response group, and their interaction, using dummy coding

confidence ~ 1 + group + illusion strength + illusion strength:group + (1 | subject)

Predictors	OR	95% CI	p.value	Cohen's d
Intercept (no-anomaly correct)	2.97	[2.65, 3.38]	**< .001**	0.60
Anomaly correct	1.07	[0.81, 1.40]	.60	0.04
Anomaly incorrect	0.64	[0.56, 0.74]	**< .001**	-0.24
Illusion strength	1.14	[1.03, 1.27]	**.005**	0.07
Anomaly correct:Illusion strength	0.79	[0.61, 1.06]	.13	-0.13
Anomaly incorrect:Illusion strength	1.02	[0.88, 1.18]	.75	0.01

	Random effects	
Group	**Parameter**	**SD**
Subject	(Intercept)	0.24

	Model fit	
Metric	**R^2 (Marginal)**	**R^2 (Conditional)**
	0.22	0.40

N	**Subject**	**Item**	**Observations**
	100	20	1,270

Note. p-values and confidence intervals were obtained using parametric bootstrap. OR = odds ratio. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

To test these predictions statistically, we used a beta generalized mixed-effects model with the initial confidence as the dependent variable (see Table 9). The fixed factors were the response group variable, the illusion strength (standardized) and their interaction, using dummy coding. The “response group” variable coded whether a given data point was a control no-anomaly trial on which the correct response was selected (intercept of the model), an anomaly trial on which the correct response was selected, or an anomaly trial on which the incorrect response was selected. Figure 5 plots the result of the regression.

Although the trends were similar to Experiment 1, the interaction term between illusion strength and the response group was not significant for correct anomaly answers[6], nor for incorrect anomaly answers. Hence, in this experiment, confidence was not significantly modulated by illusion strength.

Experiment 3

Taken together, in Experiment 2, we found that correct intuitive responding was still possible for participants despite an even more stringent deadline and a higher cognitive load than in Experiment 1. These additional constraints decreased the initial accuracy at the anomaly problems, which led to a lower non-correction rate compared to Experiment 1. Nevertheless, even with extreme constraints in Experiment 2, correct intuiting was still present. These results suggest that sound intuiting is thus an alternative route to correct responding. Concerning the error detection findings, we replicated the main results of Experiment 1. The harder deadline and load did not affect error sensitivity, suggesting the process mainly operates intuitively.

Experiment 1 suggested that illusion strength modulated error detection. When illusion strength increased, response confidence tended to decrease for correct responses and to increase for incorrect anomaly responses. In other words, error detection became less likely for “harder”

[6] However, note that the interaction term between illusion strength and correct anomaly answers reached significance in our supplementary analysis including the trials with a missed cognitive load (*p* = .04; see Supplementary Material C for the full model).

problems. However, despite similar trends, these correlational results did not reach statistical significance in Experiment 2. In Experiment 3, we therefore sought to test the impact of illusion strength experimentally, by directly manipulating the semantic overlap between the impostor (e.g., "Moses") and the undistorted original term (e.g., "Noah", Hannon & Daneman, 2001; Van Oostendorp & De Mul, 1990). This allowed us to get a more controlled measure of illusion strength. We created "weak" (easy to spot) impostor questions (e.g., "How many animals of each kind did *Goliath* take on the ark?"), compared to the "strong" (hard to spot) impostor questions we used in Experiment 1-2 (e.g., "How many animals of each kind did *Moses* take on the ark?"). In Experiment 3, participants were randomly allocated to either the weak-impostor or strong-impostor conditions.

The first goal of the experiment was to test whether the weak-impostor versions of the questions would elicit more correct intuitive responses than the strong-impostor versions of the questions. Indeed, if the correct intuition is made stronger (as we expect in the weak-impostor questions), we should observe more initial correct responses.

Second, we wanted to test whether response confidence would be modulated by the strength of the impostor in a similar fashion as in our illusion strength analysis. Compared to the strong-impostor questions, the weak-impostor questions can be expected to increase the activation of the correct intuition (e.g., "It's definitely not Goliath"). We therefore expected that participants in the weak-impostor condition would be less confident than participants in the strong-impostor condition for incorrect initial anomaly responses (i.e., intuitive error sensitivity increases for easier problems), whereas for correct responses they would be more confident than participants in the strong-impostor conditions.

In addition, in Experiment 3 we used an alternative two-block paradigm (for a similar design, see Markovits et al., 2019; Raoelison et al., 2021). We presented participants with two separate blocks of distinct trials in a randomly determined order: an intuitive block with a cognitive load task and a response deadline and a deliberative block without any cognitive load, nor response deadline.

Method

Transparency and openness.

The research question and experimental design were preregistered on the OSF platform (https://osf.io/42wr3)[7]. No specific analyses were preregistered.

Participants

We recruited 200 online participants (106 females, *M* age = 41.6 years, *SD* = 14 years) via the Prolific platform. Only native English-speaking American (USA) participants were eligible to take part in the experiment. Participants received £1 for their 10 minutes of participation. Among them, 32.5% reported high school as their highest educational level, 1% reported less than high school, and 63% reported holding a higher education degree.

[7] In the preregistration, Experiment 3 is referred to as Experiment 4 due to an additional two-response paradigm comparison experiment that we ran beforehand, which is itself labeled as Experiment 3.

Materials and Procedure

The experiment was conducted online using the OpenSesame/Osweb software (Mathôt & March, 2022) and was hosted on the freely accessible MindProbe server (https://mindprobe.eu) via the JATOS server software (Lange et al., 2015).

Semantic Similarity Manipulation. We constructed weaker versions of our anomaly problems (e.g., Hannon & Daneman, 2001). For instance, in the following anomaly question: "In the biblical story, how many animals of each kind did Moses take on the Ark?", we replaced the strong-impostor word "*Moses*" by "*Goliath*". Note that both the weak-impostor and the strong-impostor words were semantically related to the control no-anomaly target ("*Noah*"). Hence, completely unrelated words (e.g., "*Kennedy*") were avoided. However, the strong impostor was more strongly related to the control no-anomaly target than the weak impostor. For each problem, a set of possible candidate weak-impostor words were generated by the three co-authors (partly based on Hannon & Daneman, 2001). After discussion, we decided on the best alternative for each problem. In case no agreement could be reached (5 problems) the top two alternatives were both included in the pretest rating study (see below). For these problems, we selected the alternative that received the most distinctive (i.e., lowest) similarity rating in the pretest. At the start of the experiment, participants were randomly allocated to either the weak-impostor group (n = 97) or the strong-impostor group (n = 103).

For completeness, note that we also ran a supplementary two-response paradigm experiment using only weak-impostor questions, to be compared with the strong-impostor questions of Experiment 2. We make the data and text accessible on the OSF platform for the interested reader (https://osf.io/bvy3u/). However, be aware that this comparison relies on two different Prolific samples, which may cause validity issues due to random sampling differences.

Norming Study. We recruited 25 native English-speaking American (USA) participants (11 females, *M* age = 36.3 years, *SD* = 10.5 years) on the Prolific platform. For each question, participants had to first read the control, no-anomaly version of the question along with the correct answer. The strong- and the weak-impostor versions of the questions were displayed below in a random order, with the impostor words in upper case. To illustrate, here is an example of one complete trial for a given question:

> The undistorted question is: "In the biblical story how many animals of each kind did NOAH take on the Ark? (answer: Two)"
>
> How similar is each distorted sentence to the original undistorted question?
>
> Please type a number from 0 (Not at all similar) to 100 (Extremely similar) for each sentence.
>
> In the biblical story how many animals of each kind did GOLIATH take on the Ark?
>
> In the biblical story how many animals of each kind did MOSES take on the Ark?

On average, the weak-impostor versions received a significantly lower similarity rating (*M* = 23.2%, *SD* = 19.6%) than the strong-impostor versions, *M* = 38.7%, *SD* = 19.3%, *t*(24) = 6.59, *p* < .001. Furthermore, the mean rating of every individual item was higher for the weak-impostor version than for the strong-impostor version. See Table 1 for the complete list of weak-impostor questions.

Two-Block Paradigm. Participants saw two distinct blocks of questions: an intuitive block and a deliberative block. In the intuitive block, they were instructed to give their first, intuitive responses to the questions and to "give the first answer that intuitively comes to mind as quickly as possible". To make sure that deliberation was minimized in the intuitive block, we used the same deadline (5 s, with the background color shifting to yellow 1 s before the deadline) and load

(3 x 3 grid containing 4 crosses) procedures as those used in Experiment 1[8]. In the deliberative block, participants could take all the time they wanted to actively reflect on the question and were instructed to "think as deeply as possible" before providing their response. The order of the two blocks was counterbalanced among participants, with half seeing the intuitive block first and half seeing the deliberative block first.

Among our selected items, we had 20 distinct questions in terms of content, each available in two versions: an anomaly version and a no-anomaly version. Out of the 20 possible questions, 10 were presented in the anomaly version and 10 in the no-anomaly version. Importantly, a participant never encountered both the anomaly and no-anomaly versions of the same question, ensuring that each participant saw 20 unique questions in terms of content. The assignment of the questions to each version (anomaly vs. no-anomaly) was randomly determined for each participant. Each block (intuitive or deliberative) contained 5 no-anomaly and 5 anomaly questions. Crucially, to minimize the potential influence of question content on the comparison between intuitive and deliberative blocks, the assignment of questions to each block was also randomly determined for each participant.

The trivia task was introduced similarly to the previous experiments. However, the instructions regarding the two-response paradigm were replaced with the introduction of the two-block paradigm as follows:

> In the **intuitive** trials, we are interested in your **initial, intuitive response**. We want you to respond with the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible.

[8] We used the less demanding deadline and cognitive load of Experiment 1 to guarantee a sufficient number of remaining trials for analysis in the intuitive block after the exclusion of trials with failed loads or deadlines.

> To make sure that you answer as fast as possible, **a time limit was set for the intuitive trials response**, which is going to be **5 seconds**. When there is 1 second left, the background color will turn to **yellow** to let you know that the deadline is approaching. Please make sure to **answer before the deadline passes**.
>
> For the **deliberative trials**, you can **take the time to actively reflect on the question** before submitting your response. Here we want you to think as deeply as possible before you give your answer.

Throughout the entire practice session, the same two trivia questions were consistently used. Participants began by responding to two intuitive trials without any cognitive load, allowing them to familiarize themselves with the deadline procedure. Following this, they completed two load matrix trials without the trivia questions. Next, they completed two full intuitive trials with both the deadline procedure and the concurrent load. Finally, participants engaged in two deliberative trials without the deadline or cognitive load task.

At the end of the practice session, participants were informed that they had to answer 10 questions per block and that they could take a break between the two blocks. At the beginning of each block, they were informed about the type of block (i.e., intuitive or deliberative) they were about to engage in. To remind participants which block they were engaging with, the color of the answer options was green in the intuitive block and blue in the deliberative block. In addition, we added a reminder sentence under each trivia question: “Please give your intuitive answer.” and “Please give your deliberative answer.”, respectively (for the complete instructions, see Supplementary Material B).

The methodology closely resembled that of the previous experiments, with the primary difference being the adoption of a two-block paradigm instead of the two-response paradigm. In the intuitive block, the fixation cross was displayed during 2 s, followed by a 2 s display of the load matrix. Then, the trivia question appeared, with the 5 s deadline signaled by the background turning yellow after 4 s. After responding, participants indicated their confidence level before proceeding to the load recall and receiving feedback regarding their load performance. In the deliberative block, participants saw a fixation cross for 2 s, answered the trivia question without time constraints or load, and then indicated their confidence level to progress to the next trial.

To check that participants experienced time pressure in the intuitive block, we contrasted response latencies for anomaly correct responses between the intuitive and deliberative blocks (see Table S15). We built a linear mixed-effects model with response block as fixed effect using dummy coding. Results indicated that participants answered significantly faster in the intuitive block (M = 3.2 s) than in the deliberative block (M = 6.3 s), t(474.32) = 15.06, p < .001, Cohen's d = -1.59, 95% CI [-1.83, -1.36].

Exclusion Criteria. Participants did not provide their response within the deadline in 7.2% of the intuitive block trials and failed to recall the correct load pattern in 19.7% of the intuitive block trials. Consequently, 25.1% of the trials were excluded from the intuitive block, and we thus analyzed 1498 trials out of the initial 2000 intuitive trials. We did not exclude any trials from the deliberative block. On average, each participant contributed a total of 7.5 valid trials in the intuitive block (out of 10, SD = 1.8 trials, range 1—10). The exclusion rate for anomaly trials (26.3%) was slightly higher than for no-anomaly trials (23.9%). However, the exclusion rate for strong-impostor trials (26.6%) was relatively similar to the exclusion rate of weak-impostor trials (26%).

Results and Discussion

Accuracy

First, we wanted to know how correct responses in the intuitive block compared to correct responses in the deliberative block on anomaly trials. We also wanted to know whether the weak-impostor questions would elicit more correct intuitive responses than the strong-impostor questions on anomaly trials. Figure 6 displays accuracy as a function of impostor strength condition (strong vs. weak) and response block (intuitive vs. deliberative). As expected, there seemed to be no discernible accuracy difference between the strong-impostor and the weak-impostor conditions for the (identical) control no-anomaly trials. For the anomaly questions however, participants in the weak-impostor condition exhibited higher accuracy compared to those in the strong-impostor condition both in the intuitive (Strong = 13.5%; Weak = 28.9%) and deliberative (Strong = 30.1%; Weak = 47.2%) response blocks. Thus, our weak-impostor manipulation worked as intended in that it boosted performance. However, note that in a fair number of cases participants still failed to solve the weak-impostor versions correctly, even in the deliberative response block.

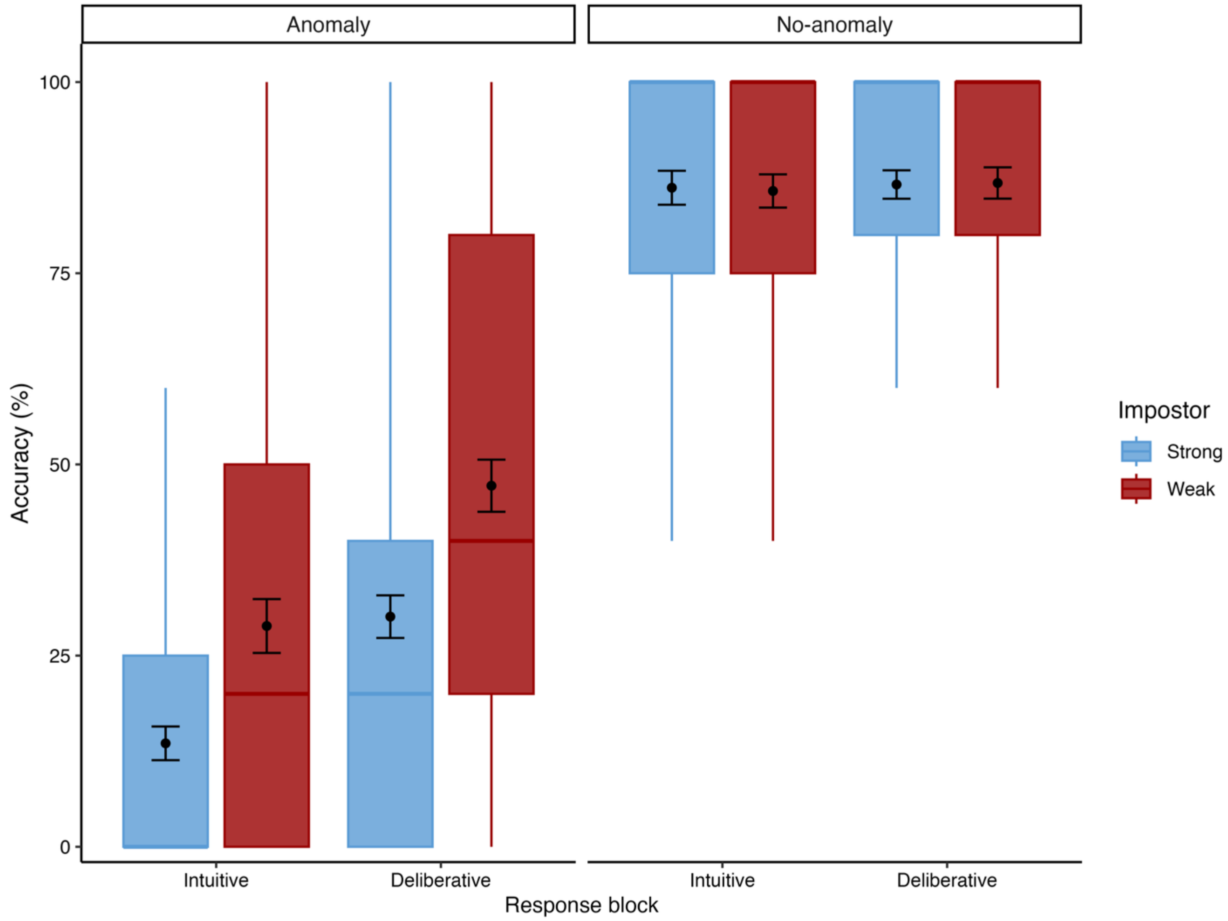

Figure 6. Response accuracy for anomaly and control no-anomaly trials as a function of response block and impostor strength in Experiment 3. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile range. Overlaid black dots represent the mean and error bars are standard errors of the mean.

To test these findings statistically, we built a binomial generalized mixed-effects model, with anomaly presence (control no-anomaly; anomaly), impostor strength (weak; strong), response block (deliberative; intuitive), and their interaction as fixed effects using sum coding (see Table 10). The main effects of impostor strength, response block, and anomaly presence were significant. Additionally, the two-way interactions between impostor strength and anomaly presence as well as between response block and anomaly presence were significant. However, the two-way interaction between impostor strength and response block as well as the three-way interaction among impostor strength, response block, and anomaly presence, were not significant. Post-hoc tests showed that for no-anomaly items, there was no significant difference between the intuitive and the deliberative response block, p = .25, or between the weak- and the strong-impostor conditions, p = .66. For anomaly items, however, both the difference between the intuitive and the deliberative response block, p < .001, and the weak- and the strong-impostor conditions were significant, p < .001. To summarize, both being in the deliberative block and in the weak-impostor condition significantly boosted participants' accuracy, but only for anomaly items.

Table 10.

Binomial generalized mixed-effects model on accuracy as a function of impostor strength, anomaly presence and response block in Experiment 3, using sum coding

accuracy ~ 1 + impostor + response block + anomaly + impostor:response block + impostor:anomaly + response block:anomaly + impostor:response block:anomaly + (1 + anomaly | subject) + (1 + anomaly | item)

Fixed effects				
Predictors	**df**	χ^2	**p.value**	**Cohen's d**
Impostor	1	10.63	**.002**	-0.19
Response Block	1	44.89	**< .001**	-0.20
Anomaly	1	46.54	**< .001**	1.05
Impostor:Response Block	1	1.09	.31	-0.03
Impostor:Anomaly	1	17.33	**< .001**	0.16
Response Block:Anomaly	1	25.69	**< .001**	0.15
Impostor:Response Block:Anomaly	1	0.17	.70	0.01

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	1.24
Item	(Intercept)	0.54
Subject	Anomaly	0.51
Item	Anomaly	0.58
Subject	Cor (Intercept x Anomaly)	-0.32
Item	Cor (Intercept x Anomaly)	0.12

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.43	0.64

N	**Subject**	**Item**	**Observations**
	200	20	3,498

Note. p-values were obtained through bootstrapping on the comparison between a full model and a reduced model without the variable of interest, using type III sum of squares.

Once again, the high accuracy (*M* = 86%) on the control, no-anomaly items in the intuitive response block rules out a potential guessing confound in the intuitive response block. The nature of errors made in anomaly problems during the intuitive response block – representing 79% of the total number of responses for the anomaly problems in the intuitive block - further supports this conclusion: the majority stemmed from giving the undistorted answer (76.8%), rather than the filler answer (15.8%) or the "Don't know" response option (7.4%).

Note that given the two-block design using different items for the intuitive and the deliberative response block, we could not compute a direction of change index per se. However, to get an indication of the prevalence of sound intuiting, we can compare the rate of correct responses in the intuitive and the deliberative blocks at the item level (i.e., correct intuitive accuracy/correct deliberate accuracy). Similar to Experiment 1-2, we observed a high prevalence of sound intuiting, with a slightly higher occurrence for weak- (*M* = 66.3%) compared to strong-impostor items (*M* = 61.9%).

In sum, participants were thus able to generate correct responses in the intuitive response block where deliberation was minimized, even in the strong-impostor condition, replicating the results of the previous experiments. Moreover, participants in the weak-impostor condition had a significantly higher accuracy compared to those in the strong-impostor condition.

Confidence Ratings

Impostor Strength Manipulation. We expected that participants in the weak-impostor condition would demonstrate lower intuitive confidence compared to participants in the strong-impostor condition for incorrect anomaly responses (i.e., suggesting an increase in intuitive error sensitivity for easier problems)[9].

[9] Following our preregistration, we did not necessarily expect that participants in the weak-impostor condition would display a higher intuitive confidence than participants in the strong-impostor conditions for correct anomaly responses due to the possibility of a ceiling effect on correct responses.

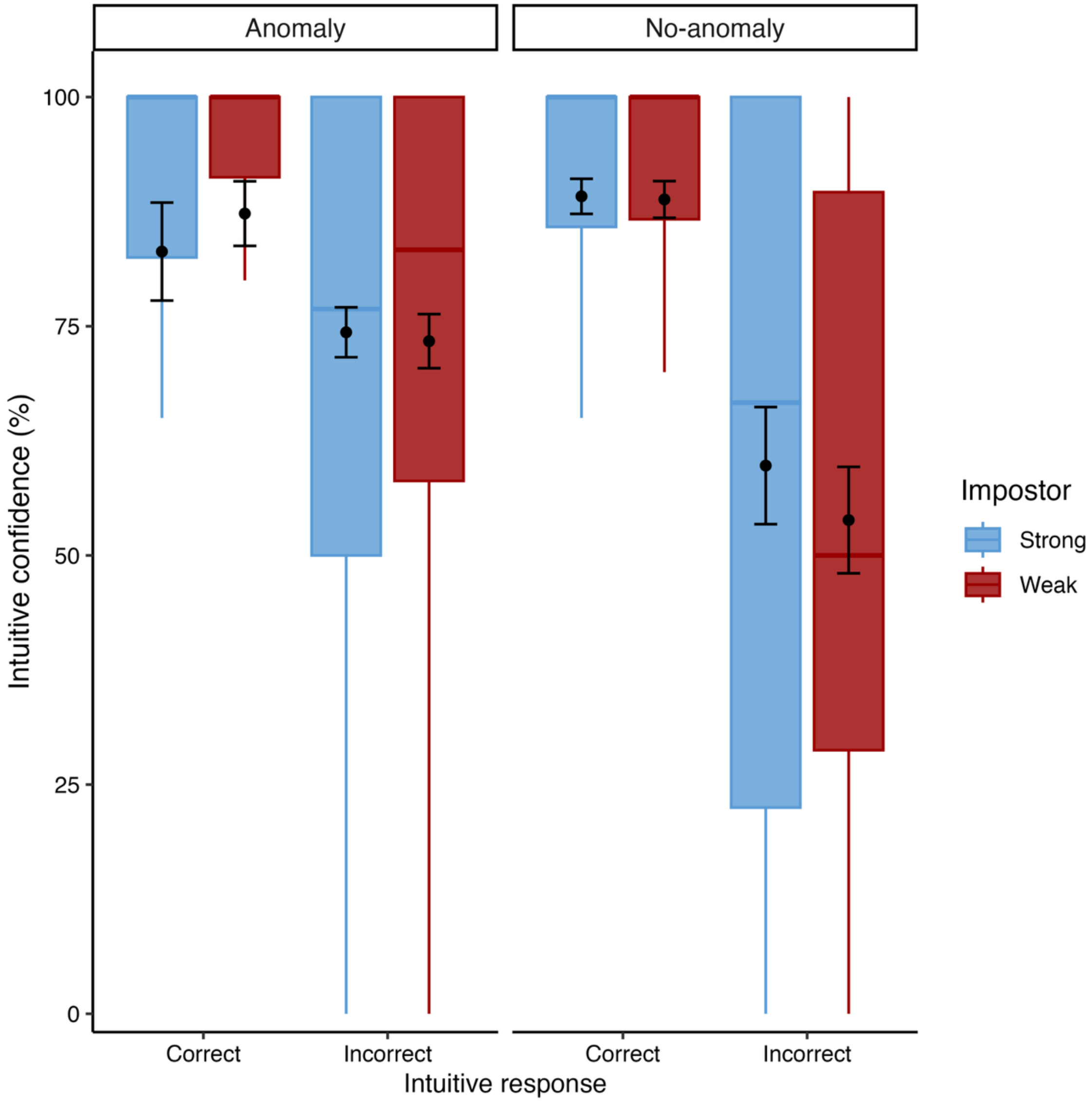

Figure 7. Confidence at anomaly and control no-anomaly trials as a function of impostor strength in the intuitive response block. The lower and upper hinges of the boxplot correspond to the first and third quartiles, and the middle line shows the median. The lower (resp. upper) whiskers extend from the hinges to the smallest (resp. largest) value no further than 1.5 times the interquartile range. Overlaid black dots represent the mean and error bars are standard errors of the mean.

Figure 7 shows the confidence ratings in the intuitive response block as a function of impostor strength. For correct control no-anomaly responses, confidence was consistently high, regardless of impostor strength (Strong = 89.2%; Weak = 88.8%). However, for incorrect anomaly responses, confidence ratings were significantly lower than for correct no-anomaly responses, replicating the error sensitivity effect observed in previous experiments, though impostor strength did not seem to modulate it (Strong = 74.3%; Weak = 73.4%). Similarly, for correct anomaly responses, confidence seemed unaffected by the impostor strength variable (Strong = 83.1%; Weak = 87.3%). Finally, for incorrect control no-anomaly responses, confidence also seemed unaffected by the impostor strength condition (Strong = 59.8%; Weak = 53.9%).

Table 11

Mixed-effects beta-regression on confidence as a function of response group and impostor strength in Experiment 3, using dummy coding

confidence ~ 1 + impostor + group + impostor:group + (1 | subject) + (1 | item)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly correct impostor strong)	4.15	[3.50, 4.85]	**< .001**	0.78
Impostor weak	0.99	[0.80, 1.22]	.93	-0.01
Anomaly correct	1.00	[0.71, 1.38]	.99	0.00
Anomaly incorrect	0.64	[0.54, 0.78]	**< .001**	-0.24
Anomaly correct:Impostor weak	1.09	[0.71, 1.68]	.71	0.05
Anomaly incorrect:Impostor weak	0.90	[0.69, 1.18]	.46	-0.06

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.37
Item	(Intercept)	0.13

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.19	0.63

N	Subject	Item	Observations
	199	20	1,388

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

To test these results statistically, we used a beta generalized mixed-effects model using confidence in the intuitive response block as our dependent variable (see Table 11). As fixed effects, we included impostor strength, “response group”, and their interaction. The response group variable encoded whether a specific data point was a control trial where the correct response was chosen (intercept of the model), an anomaly trial where the correct response was chosen, or an anomaly trial where the incorrect response was chosen. Both impostor strength and response group were dummy coded variables. The only significant term in the model was the coefficient for anomaly incorrect response, indicating that participants exhibited a significant error sensitivity effect. However, none of the interaction terms between impostor strength and anomaly correct responses, or between impostor strength and anomaly incorrect responses were significant.

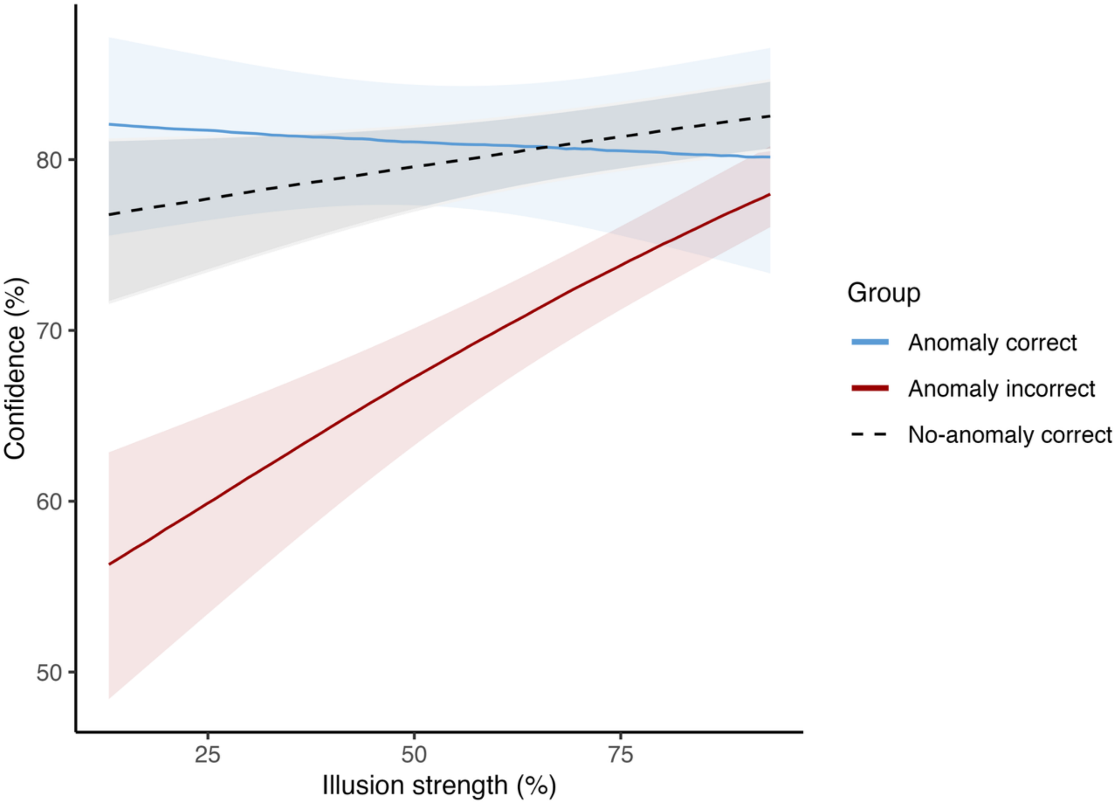

Figure 8. Regression results of initial confidence as a function of illusion strength for control no-anomaly correct (baseline), anomaly correct and anomaly incorrect responses in Experiment 3. Illusion strength = mean initial no-anomaly accuracy − mean initial anomaly accuracy for each item. The shaded bands are 95% confidence bands.

Illusion Strength Analysis. For completeness, at the item level, we also performed an additional illusion strength analysis similar to that in Experiment 1 and 2 (see Table 12). For each item, we computed the illusion strength (i.e., how difficult it is to spot the illusion) by subtracting the mean intuitive accuracy of the anomaly version from the mean intuitive accuracy of the control no-anomaly version. A higher difference between the two indicates a stronger illusion.

Table 12

Mixed-effects beta-regression on confidence as a function of response group and standardized illusion strength in Experiment 3, using dummy coding

confidence ~ 1 + group + illusion strength + illusion strength:group + (1 | subject)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly correct)	4.16	[3.71, 4.69]	**< .001**	0.79
Anomaly correct	1.01	[0.78, 1.27]	.95	0.00
Anomaly incorrect	0.59	[0.51, 0.67]	**< .001**	-0.29
Illusion strength	1.09	[0.99, 1.21]	.07	0.05
Anomaly correct:Illusion strength	0.89	[0.71, 1.11]	.29	-0.07
Anomaly incorrect:Illusion strength	1.18	[1.02, 1.37]	**.032**	0.09

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.37

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.26	0.64

N	**Subject**	**Item**	**Observations**
	199	40	1,388

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

We used a beta generalized mixed-effects model with the initial confidence as the dependent variable. The fixed factors were the response group variable, the illusion strength and their interaction, using dummy coding. The “response group” variable coded whether a given data point was a control no-anomaly trial on which the correct response was selected (intercept of the model), an anomaly trial on which the correct response was selected, or an anomaly trial on which the incorrect response was selected. Figure 8 plots the result of the regression. Critically, the interaction term between illusion strength and the response groups was significant for incorrect anomaly answers[10], but not for correct anomaly answers. Hence, as illusion strength increased (i.e., the alleged correct intuition decreased), confidence increased for incorrect anomaly answers, indicating that error sensitivity became less pronounced for “harder” problems, whereas confidence was not significantly modulated by illusion strength for correct anomaly answers.

General Discussion

The present experiments tested the dual process view of semantic illusions. According to this account, giving the correct answer to an anomaly question requires deliberate processing to overcome an intuitively cued incorrect answer. By adopting a two-response paradigm in which participants had to give an initial response under time pressure and a concurrent cognitive load, we first tested whether the correct response could also be generated intuitively. Across our experiments, we found that participants were often able to generate correct intuitive responses to anomaly questions. Our results thus suggest that sound intuiting is a non-negligible alternative route to correct responding in the case of semantic illusions.

Second, the dual process account of semantic illusions further assumes that falling for the illusion results from a failure to engage in deliberate processing and to detect the anomaly in the

[10] The interaction term between illusion strength and anomaly incorrect answers ceased to be significant upon excluding the “Don’t know” responses (*p* = .08). This result should thus be interpreted with caution.

sentence. In each of our three experiments, we consistently observed that participants who gave incorrect responses to an anomaly question were sensitive to the erroneous nature of their response (as reflected in a decreased confidence). Critically, this error detection effect was present at the initial, intuitive stage, suggesting that this error sensitivity is an automatic and effortless process.

Third, the illusion strength analyses of Experiment 1, 2 and 3 and the impostor strength manipulation of Experiment 3 suggest that both confidence and performance in the initial, intuitive stage depend on the strength of the illusion. When the illusion was weaker, sound intuiting became more frequent and confidence tended to decrease in the case of an incorrect response (i.e., more error sensitivity). However, it should be noted that the impact of illusion strength and impostor strength on confidence did not always reach significance, suggesting these results should be approached with some caution (see Supplementary Material A for our sensitivity power analyses).

These results have both theoretical and practical implications. At the theoretical level, our results allow us to better understand the nature of erroneous and correct responding in semantic illusions. Our error sensitivity findings indicate that errors do not always result from a failure to detect the anomaly in the distorted sentence. Hence, when responding incorrectly to the question "In the biblical story, how many animals of each kind did Moses take on the Ark?", a substantial number of participants must also activate the correct concept of "Noah" on some level (see Supplementary Material F). In addition, our results show that correct responding does not necessarily require deliberation but that it can frequently be achieved intuitively. Overall, these findings are consistent with recent advances in dual process theorizing in which the intuitive performance to classic "bias" tasks is determined by the interplay of both incorrect and correct intuitions (De Neys, 2023). According to this dual process model 2.0. (De Neys, 2017), the logical, correct response that has traditionally been considered to be cued by deliberation, can also be cued intuitively through System 1. Hence, the model assumes that System 1 will cue both an incorrect "heuristic" and correct "logical" intuition in a typical heuristics-and-biases task. Whichever intuition is strongest will be selected as initial response. In addition, the more similar the strength of the competing intuitions, the more conflict will be experienced, and the more one will doubt their decision. The current findings are in line with this view, as they also suggest that responders faced with semantic illusions automatically generate both a correct ("Noah") and an incorrect ("Two") intuition. The illusion strength analyses and the impostor strength manipulation tentatively support the idea that the strength interplay of these intuitions may determine the intuitive performance in semantic illusions (i.e., how likely it is one responds correctly and how likely it is one shows error detection).

Our findings may also inform language comprehension research. As noted in the introduction, Park and Reder (2004) argued that "distortion detection involves a two-pass process—the first to flag a potential mismatch and the second to invoke a careful inspection that might confirm an erroneous term in the question" (p. 282)." According to this view, participants initially rely on a "partial matching" mechanism, engaging in a careful examination of the distorted sentence only if the semantic overlap between the memory trace and the presented sentence is sufficiently low. This view is supported by psycholinguistic research on so-called "garden-path" sentences, where an ambiguous sentence is resolved in an unexpected manner (e.g., "The old man the boats", where we first interpret the word "old" as an adjective instead of a noun, and the word "man" as a noun instead of a verb). Indeed, people tend to reevaluate their initial sentence processing decisions when they encounter semantic or syntactic incoherence (e.g., Blott et al., 2021; Ferreira et al., 2001). Our confidence results suggest that this reevaluation may occur through an intuitive metacognitive process, prompting the revision of the initial incorrect interpretation. However, our results also suggest that the "careful inspection", which Park and Reder describe as following the initial mismatch detection, can also be completed intuitively. This intuitive detection of semantic anomalies we observed throughout our experiments is supported by research showing that experienced language users often arrive at the correct interpretation of complex sentences from the start, without needing to deliberately correct an initial misunderstanding (Ferreira & Huettig, 2023). This may be explained by the automation of initially challenging linguistic forms through practice (Favier et al., 2021).

Note that our claim regarding error sensitivity pertains solely to the relative difference in confidence between correct no-anomaly answers and incorrect anomaly answers. We do not argue that participants were well-calibrated in this task, given their consistently high confidence ratings on anomaly incorrect answers. Nevertheless, their confidence ratings still allowed to discriminate between correct and incorrect responses, following the definition of metacognitive sensitivity (Ackerman & Thompson, 2017; Fleming & Lau, 2014). However, note that the metacognitive sensitivity for incorrect anomaly responses was still lower than for the (rare) incorrect no-anomaly responses —where participants mainly gave “Don’t know” or filler responses (e.g., “Three” in the original Moses question). Participants were thus less confident when responding incorrectly to an undistorted question than when failing to notice the illusion in a distorted question. This likely reflects different processes: in no-anomaly incorrect responses, confidence is very low because people clearly know they don’t know the answer. In incorrect anomaly responses, confidence is higher, indicating that here error sensitivity reflects a more subtle process that does not necessarily equate with conscious error detection per se (for a similar point, see Mata, 2023). This high confidence in incorrect anomaly answers may be well accounted for by self-consistency models of confidence, according to which confidence is based on the agreement among accessed representations supporting the current answer, thus reflecting the probability of making the same choice on the repetition of the current problem (e.g., Koriat, 2012).

These results also bear practical implications, as semantic illusions have been used as a measure of people's capacity and disposition to engage in effortful deliberation (e.g., Sirota et al., 2021). However, our results indicate that correct answers in the case of semantic illusions are often generated intuitively. Therefore, the mere use of correct answers on semantic illusion problems as an index of deliberate processing abilities can be problematic and distort conclusions. To clearly measure one process or the other (i.e., intuition or deliberation capacities), we recommend testing how the answer has been generated (e.g., by using a two-response paradigm).

Critics of our work may argue that we observed correct intuiting because the items we selected might have been relatively easy compared to those typically used in the literature. To test this hypothesis, we can directly compare our illusion rate to the results in the literature. Following Speckmann and Unkelbach (2021), we computed the rate of incorrect undistorted responses (e.g., "Two") at anomaly questions in the final response stage across Experiment 1 and Experiment 2 (i.e., the % of trials in which participants fell prey to the illusion). The result (56%) was higher than what Reder and Kusbit (1991) reported (33% in Experiment 1, 35% in Experiments 2 & 3, 32% in Experiment 4; we only used the results from the comparable literal task condition). Similarly, Speckmann and Unkelbach (2021) also found lower rates than the ones reported here (49% in Experiment 1, 52.6% in Experiment 2). This slightly higher illusion rate in our experiment may be explained by the fact that we only selected items from Speckmann and Unkelbach (2021) which had a high knowledge check as well as a high control no-anomaly accuracy. Hence, if anything, our items were overall harder than those adopted in the literature which implies that sound intuiting will be even more prevalent in other studies.

Another critique might be that correct intuiting was possible in our experiments because our design was not challenging enough and still allowed deliberation. To minimize the possibility that participants engage in deliberate processing in the initial stage, we combined 3 validated procedures (instructions, time pressure, and concurrent load). All these manipulations have been previously shown to minimize deliberation (Bago & De Neys, 2017). In Experiments 2, we used an even more challenging load task and deadline to further minimize the possibility that participants would deliberate in the initial response stage. Nevertheless, one may still argue that we could have used an even more demanding deadline and load task. However, the high number of missed trials in Experiment 2 (31.6%) shows that adding load or time pressure would have raised practical and statistical issues (i.e., selection of participants due to a large portion of discarded trials, e.g., Bouwmeester et al., 2017). From a more theoretical standpoint, the problem is that dual process theories are underspecified (Kruglanski, 2013). The framework often entails that System 2 is slower and more demanding than System 1 but gives us no unequivocal a priori criterion that allows us to classify a process as intuitive or deliberate (e.g., takes at least x time, or x amount of load; De Neys, 2023). Consequently, as long as we keep on observing correct initial responses, one can always argue that these will disappear “with just a little bit more load/time pressure”. However, note that the corrective assumption becomes unfalsifiable at this point. Any evidence for correct intuiting can always be explained away by arguing that the methodological design let room for deliberation. At the same time, we acknowledge that we cannot claim that all possible deliberation was ruled out in our studies.

Although the conflict detection effect we observed is in line with findings in the logical reasoning field (e.g., Bago & De Neys, 2017, 2020), sound intuiting was less prevalent in our semantic illusion task than what is typically found in the reasoning field—where correct intuitive responding tends to be the modal pattern over deliberative correction. For instance, across four experiments, Bago and De Neys (2017) reported higher “non-correction rates” for syllogistic reasoning (*M* = 87.6%) and base-rates (*M* = 74.8%) tasks than what we found (46.1% in

Experiment 1; 28.1% in Experiment 2)[11]. We speculate that this may be explained by the different nature of our task. Classic reasoning tasks can be solved using a universal algorithm. Once you know the correct rule, you can apply it whatever the specific values in the problem are. For instance, in a base-rate task, you simply have to give weight to the priors/base-rates that are given in the problem. Likewise, in the bat-and-ball problem one might use the equation "x + y = a. x = y + b. Solve for x", for example. The solution strategy can thus be automatized, which is assumed to be the nature of correct intuitions in these tasks (De Neys, 2012; Raoelison et al., 2020). However, in the case of semantic illusions, there is no general algorithm one could apply. Instead, one can only carefully search their semantic memory, but this search will be "unique" for each problem. Hence, this semantic search strategy might be less automatized than applying the correct rule in a reasoning problem. Therefore, "correct" responses might be less instantiated than in classic reasoning tasks which would explain the lower prevalence of correct intuiting in the case of semantic illusions.

To conclude, we believe that it is hard for the dual process account of semantic illusions to account for our findings, and that they rather support recent models in which the absolute and relative strength of competing intuitions determines performance.

Acknowledgements

This research was supported by a research grant (DIAGNOR, ANR16-CE28-0010-01) from the Agence Nationale de la Recherche, France.

[11] Experiments using moral (Bago & De Neys, 2019a) and prosocial (Bago et al., 2021) reasoning tasks also yielded very high overall "non-correction rates" (83.8% and 83.1% respectively).

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bago, B., Bonnefon, J.-F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*. 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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4* (arXiv:1406.5823). arXiv. http://arxiv.org/abs/1406.5823

Beucler, J., Voudouri, A., & De Neys, W. (2021, November 8). Intuitive Responding in the Moses Illusion. https://doi.org/10.17605/OSF.IO/BPMC8

Beucler, J., Voudouri, A., & De Neys, W. (2021, December 16). Intuitive Responding in the Moses Illusion - Experiment 2. https://doi.org/10.17605/OSF.IO/295BF

Beucler, J., Voudouri, A., & De Neys, W. (2022, November 4). Intuitive Responding in the Moses Illusion - Experiment 4. https://doi.org/10.17605/OSF.IO/42WR3

Beucler, J., Voudouri, A., & De Neys, W. (2024, February 9). A two-response paradigm investigation of the Moses illusion. Retrieved from https://osf.io/bvy3u/

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

Blott, L. M., Rodd, J. M., Ferreira, F., & Warren, J. E. (2021). Recovery from misinterpretations during online sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(6), 968–997. https://doi.org/10.1037/xlm0000936

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bottoms, H. C., Eslick, A. N., & Marsh, E. J. (2010). Memory and the Moses illusion: Failures to detect contradictions with stored knowledge yield negative memorial consequences. *Memory*, *18*(6), 670–678. https://doi.org/10.1080/09658211.2010.501558

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G., Cornelissen, G., Døssing, F. S., & Espín, A. M. (2017). Registered replication report: Rand, greene, and nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. https://doi.org/10.1177/1745691617693624

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.3929/ethz-b-000240890

Burič, R., & Konrádová, L. (2021). Mindware instantiation as a predictor of logical intuitions in the Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology*, *32*(4), 460–477. https://doi.org/10.1080/20445911.2020.1766472

Büttner, A. C. (2012). The effect of working memory load on semantic illusions: What the phonological loop and central executive have to contribute. *Memory*, *20*(8), 882–890. https://doi.org/10.1080/09658211.2012.706308

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Cantor, A. D., & Marsh, E. J. (2017). Expertise effects in the Moses illusion: Detecting contradictions with stored knowledge. *Memory*, *25*(2), 220–230. https://doi.org/10.1080/09658211.2016.1152377

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 817–828. https://doi.org/10.1080/17470218.2015.1134603

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. (2017). Bias, conflict, and fast logic: Towards a hybrid dual process future? In *Dual process theory 2.0* (pp. 47–65). Routledge. https://doi.org/10.4324/9781315204550

De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on Psychological Science*, *16*(6), 1412–1427. https://doi.org/10.1177/1745691620964172

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68. https://doi.org/10.1017/S0140525X2200142X

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

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., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, *54*(2), 128–133. https://doi.org/10.1027/1618-3169.54.2.128

De Neys, W., & Verschueren, N. (2006). Working memory capacity and a notorious brain teaser: The case of the Monty Hall Dilemma. *Experimental Psychology*, *53*(2), 123–131. https://doi.org/10.1027/1618-3169.53.1.123

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551. https://doi.org/10.1016/S0022-5371(81)90165-1

Evans, J. S. B. (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., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Favier, S., Meyer, A. S., & Huettig, F. (2021). Literacy can enhance syntactic prediction in spoken language processing. *Journal of experimental psychology. General*, *150*(10), 2167–2174. https://doi.org/10.1037/xge0001042

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: implications for models of sentence processing and reanalysis. *Journal of psycholinguistic research*, *30*(1), 3–20. https://doi.org/10.1023/a:1005290706460

Ferreira, F., & Huettig, F. (2023). Fast and slow language processing: A window into dual-process models of cognition. *Behavioral and Brain Sciences*, *46*, e121. https://doi.org/10.1017/S0140525X22003041

Ferreira, F., & Patson, N. D. (2007). The ‘Good Enough’ Approach to Language Comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Hannon, B., & Daneman, M. (2001). Susceptibility to semantic illusions: An individual-differences perspective. *Memory & Cognition*, *29*(3), 449–461. https://doi.org/10.3758/BF03196396

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64. https://doi.org/10.1016/j.actpsy.2015.12.008

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

Kamas, E. N., Reder, I. M., & Ayers, M. S. (1996). Partial matching in the Moses illusion: Response bias not sensitivity. *Memory & Cognition*, *24*(6), 687–699. https://doi.org/10.3758/BF03201094

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 1013–1040. https://doi.org/10.1080/17470218.2015.1053951

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80. https://doi.org/10.1037/a0025648

Koriat, A. (2017). Can people identify “deceptive” or “misleading” items that tend to produce mostly wrong answers? *Journal of Behavioral Decision Making*, *30*(5), 1066–1077. https://doi.org/10.1002/bdm.2024

Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 242–247. https://doi.org/10.1177/1745691613483477

Lakens, D. (2022). Sample size justification. *Collabra: psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lange, K., Kühn, S., & Filevich, E. (2015). " Just another tool for online studies”(JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *Package “emmeans”*.

Low, J., Butterfill, S. A., & Michael, J. (2023). A view from mindreading on fast-and-slow thinking. *The Behavioral and brain sciences*, *46*, e130. https://doi.org/10.1017/S0140525X22002898

Lüdecke, D., Ben-Shachar, M., Patil, I., & Makowski, D. (2020). Extracting, Computing and Exploring the Parameters of Statistical Models using R. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445

March, D. S., Olson, M. A., & Gaertner, L. (2023). Automatic threat processing shows evidence of exclusivity. *The Behavioral and brain sciences*, *46*, e131. https://doi.org/10.1017/S0140525X22002928

Markovits, H., de Chantal, P. L., Brisson, J., & Gagnon-St-Pierre, É. (2019). The development of fast and slow inferential responding: Evidence for a parallel development of rule-based and belief-based intuitions. *Memory & cognition*, *47*, 1188-1200. https://doi.org/10.3758/s13421-019-00927-3

Mata, A. (2023). Overconfidence in the Cognitive Reflection Test: Comparing Confidence Resolution for Reasoning vs. General Knowledge. *Journal of Intelligence*, *11*(5), 81. https://doi.org/10.3390/jintelligence11050081

Mata, A., Ferreira, M. B., & Reis, J. (2013). A process-dissociation analysis of semantic illusions. *Acta Psychologica*, *144*(2), 433–443. https://doi.org/10.1016/j.actpsy.2013.08.001

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: an attentional account of reasoning errors. *Psychonomic bulletin & review, 24*(6), 1980–1986. https://doi.org/10.3758/s13423-017-1234-7

Mata, A., Schubert, A. L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: how "good-enough" representations induce biases. *Cognition, 133*(2), 457–463. https://doi.org/10.1016/j.cognition.2014.07.011

Mathôt, S., & March, J. (2022). Conducting Linguistic Experiments Online With OpenSesame and OSWeb. *Language Learning*, *72*(4), 1017–1048. https://doi.org/10.1111/lang.12509

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621. https://doi.org/10.1037/0096-3445.130.4.621

Park, H., & Reder, L. (2004). *Moses Illusion: Implication for Human Cognition*. https://doi.org/10.1184/R1/6617207.V1

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

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

Raoelison, M., Keime, M., & De Neys, W. (2021). Think slow, then fast: Does repeated deliberation boost correct intuitive responding? *Memory & Cognition*, *49*(5), 873–883. https://doi.org/10.3758/s13421-021-01140-x

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

Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory and Language*, *30*(4), 385–406. https://doi.org/10.1016/0749-596X(91)90013-A

Shafto, M., & MacKay, D. G. (2000). The Moses, mega-Moses, and Armstrong illusions: Integrating language comprehension and semantic memory. *Psychological Science*, *11*(5), 372–378. https://doi.org/10.1111/1467-9280.00273

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2015). Afex: Analysis of factorial experiments. *R Package Version 0.13–145*.

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

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54. https://doi.org/10.1037/1082-989X.11.1.54

Speckmann, F., & Unkelbach, C. (2021). Moses, money, and multiple-choice: The Moses illusion in a multiple-choice format with high incentives. *Memory & Cognition*, *49*(4), 843–862. https://doi.org/10.3758/s13421-020-01128-z

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

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

Stupple, E. J., & Ball, L. J. (2008). Belief–logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning, 14*(2), 168-181. https://doi.org/10.1080/13546780701739782

Stupple, E. J., Ball, L. J., Evans, J. S. B., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology, 23*(8), 931-941. https://doi.org/10.1080/20445911.2011.589381

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. B. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General*, *147*(7), 945. https://doi.org/10.1037/xge0000457

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

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

Trémolière, B., De Neys, W., & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, *124*(3), 379–384. https://doi.org/10.1016/j.cognition.2012.05.011

Van Oostendorp, H., & De Mul, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta Psychologica*, *74*(1), 35–46. https://doi.org/10.1016/0001-6918(90)90033-C

Verkuilen, J., & Smithson, M. (2012). Mixed and Mixture Regression Models for Continuous Bounded Responses Using the Beta Distribution. *Journal of Educational and Behavioral Statistics*, *37*(1), 82–113. https://doi.org/10.3102/1076998610396895

Verschueren, N., Schaeken, W., & d'Ydewall, G. (2004). Everyday conditional reasoning with working memory preload. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*.

Voeten, C. C. (2020). buildmer: Stepwise elimination and term reordering for mixed-effects regression. *R Package Version*, *1*(6).

Voudouri, A., Białek, M., Domurat, A., Kowal, M., & De Neys, W. (2022). Conflict detection predicts the temporal stability of intuitive and deliberate reasoning. *Thinking & Reasoning*, 1-29. https://doi.org/10.1080/13546783.2022.2077439

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Zeileis, A., Cribari-Neto, F., Gruen, B., Kosmidis, I., Simas, A. B., Rocha, A. V., & Zeileis, M. A. (2016). Package 'betareg.' *R Package*, *3*(2).

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, *95*, 1–36. https://doi.org/10.18637/jss.v095.i01

Supplementary Material

A. Sensitivity power analyses

To estimate the detectable effect size based on our data, we conducted sensitivity power analyses. These analyses assess whether our design had adequate power to identify various effect sizes during the hypothesis testing process (Lakens, 2022). Sensitivity power analyses are deemed more adequate than post hoc power analysis based on the observed effect size, because the latter are directly related to the *p*-value of the statistical test of the study (Hoenig & Heisey, 2001) and are thus uninformative.

We used the SIMR package (Green & MacLeod, 2016) to simulate the data from our models. Note that because the beta generalized mixed-effects model was unsupported for running the simulations, we refitted our beta-regression models using linear mixed-effects models on confidence (range 0-100%).

We performed four distinct sensitivity power analyses based on our main analyses in our three experiments. We conducted Monte Carlo simulations to estimate the power across a range of plausible effect sizes in our mixed-models (Kumle et al., 2021). For each analysis, we simulated new datasets with various effect sizes and tested each one for statistical significance. Since we used Wald Z-tests throughout our simulations for computational efficiency, we set a significance threshold $\alpha = .045$ to compensate for the anti-conservative asymptotic approximation. Subsequently, we calculated power as the proportion of significant results across our simulations for each effect size. Importantly, we used our data and estimated models as a starting point for these simulations, directly manipulating the effect sizes for each simulation to establish the true effect sizes from which to compute power. This approach allowed us to consider crucial information in our simulations, such as random structure or excluded trials.

Accuracy (Experiment 1)

We used the data (n = 100) from Experiment 1, along with the binomial generalized mixed-effects model on accuracy as a function of anomaly presence and response stage reported in Table 2. The statistical test from which we computed power, focused on one fixed effect in the model, for which we used a plausible range of effect sizes, ranging from Cohen's *d* = 0.20 to 0.5 using steps of 0.05. As showed in Figure S1, our simulations show that we were able to detect a Cohen's *d* of 0.35 with 80% power.

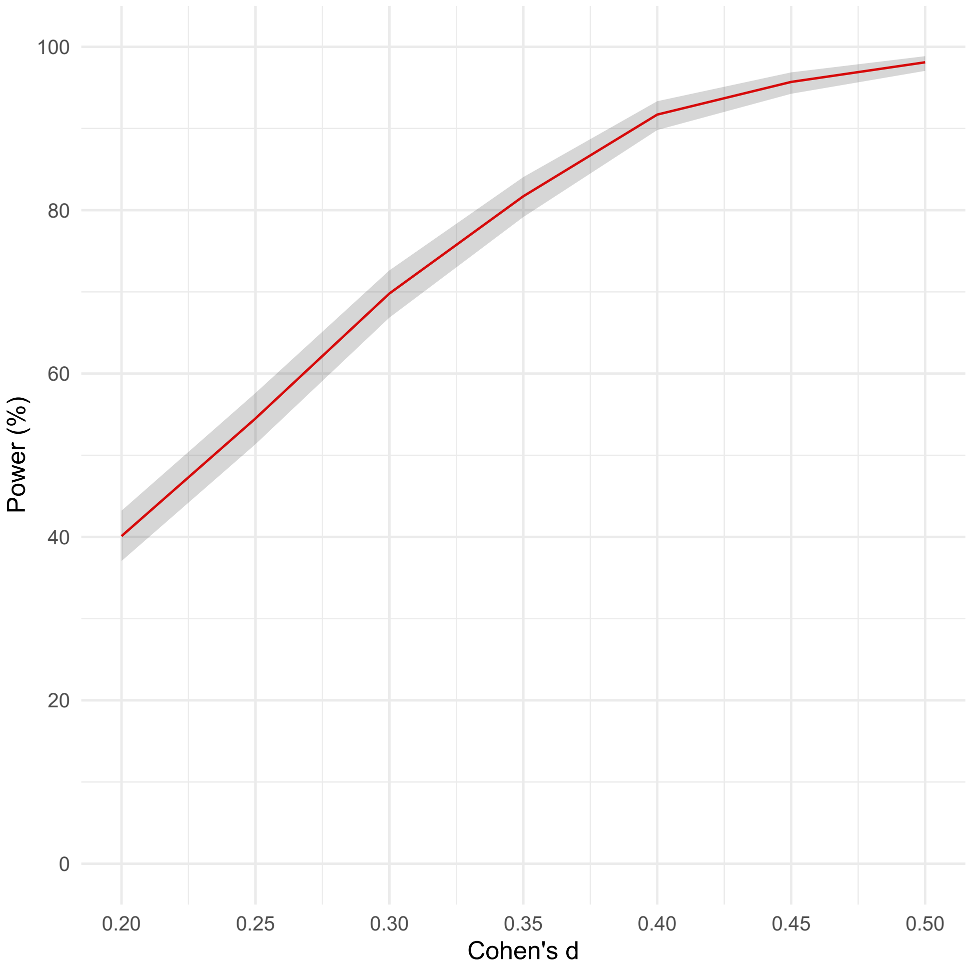

Figure S1. Sensitivity power curve illustrating the achieved power as a function of effect size for n=100 on accuracy in Experiment 1, based on 1000 simulations per effect size.

Accuracy (Experiment 3)

We used the data (n = 200) from Experiment 3, along with the binomial generalized mixed-effects model on accuracy as a function of impostor strength, response block and anomaly presence reported in Table 10. The statistical test from which we computed power focused on one fixed effect in the model, for which we used a plausible range of effect sizes ranging from Cohen's *d* = 0.20 to 0.5 using steps of 0.05. As showed in Figure S2, our simulations show that we were able to detect a Cohen's *d* of less than 0.10 with 80% power.

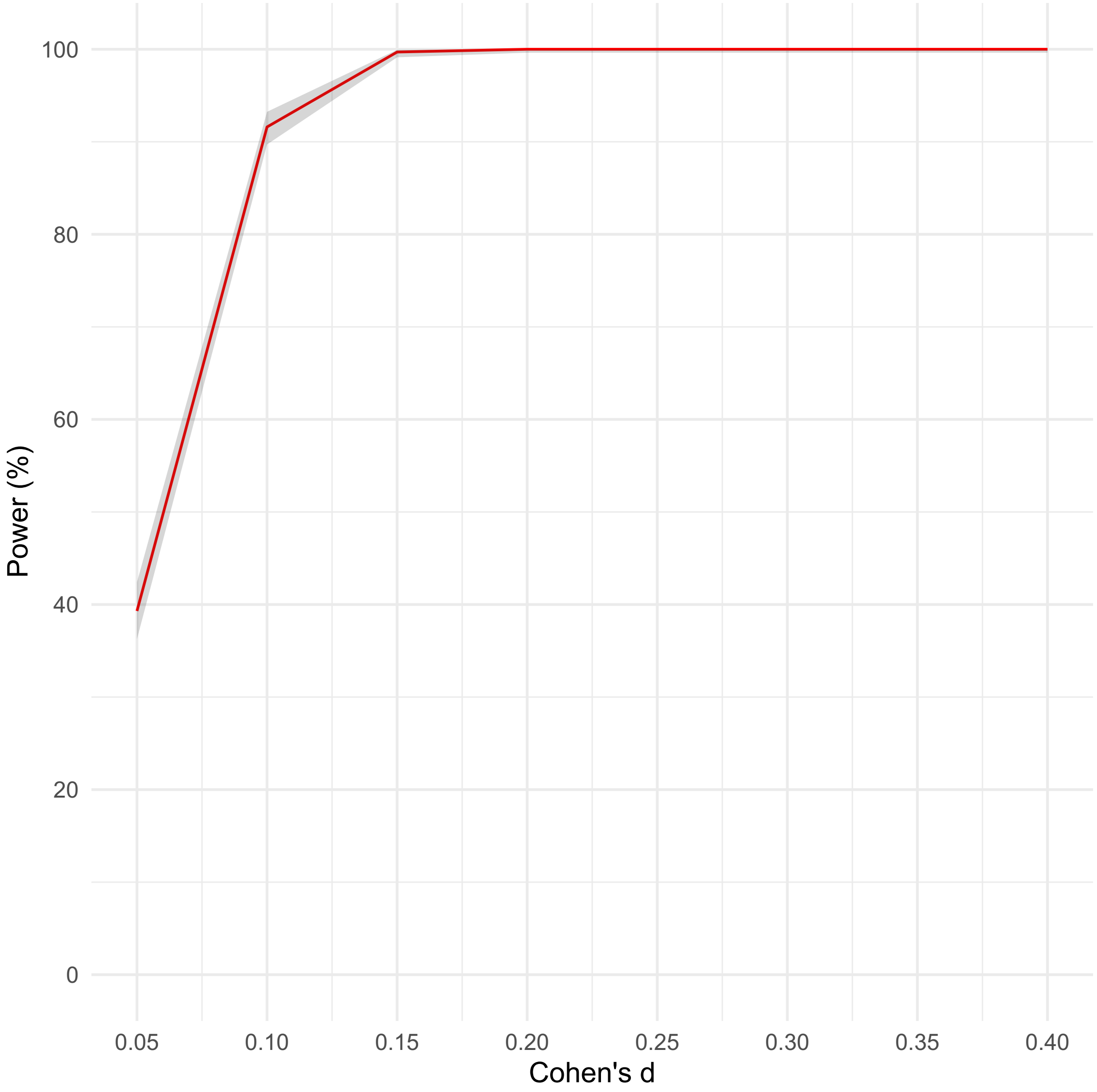

Figure S2. Sensitivity power curve illustrating the achieved power as a function of effect size for n=200 on accuracy in Experiment 3, based on 1000 simulations per effect size.

Confidence (Experiment 1)

We used the data (n=100) from Experiment 1, along with a linear mixed-effects model on confidence using the same formula and random structure as the beta generalized mixed-effects model reported in Table 3. We focused on the fixed effect interaction term contrasting the no-anomaly correct answers to the anomaly incorrect answers. We computed power for a plausible range of effect sizes for this error sensitivity effect, ranging from 1 to 8 percentage points of confidence. As shown in Figure S3, our simulations indicate that we were able to detect a 4.6 percentage point difference in confidence with 80% power.

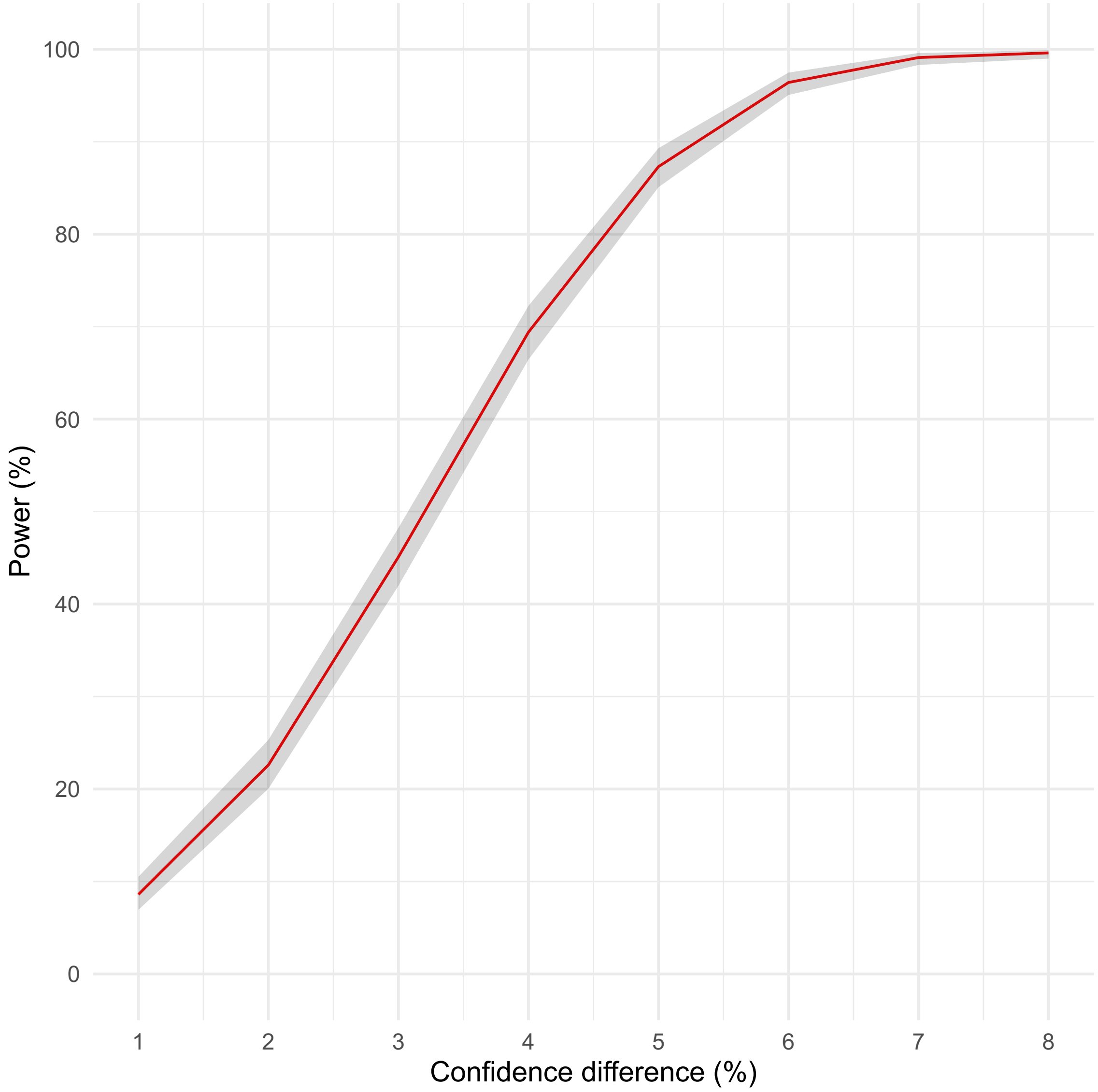

Figure S3. Sensitivity power curve illustrating the achieved power as a function of effect size for n=100 on confidence in Experiment 1, based on 1000 simulations per effect size. Confidence difference denotes the magnitude of the contrast between no-anomaly correct and anomaly incorrect responses in confidence percentage points.

Confidence (Experiment 3)

We used the data (n=200) from Experiment 3, along with a linear mixed-effects model on confidence using the same formula and random structure as the beta generalized mixed-effects model reported in Table 11. We focused on the interaction term between incorrect anomaly responses and impostor strength (testing whether error sensitivity was stronger for weak- compared to strong-impostor questions. We directly manipulated the size of this interaction, selecting a plausible range of values from a 1% to 10% confidence difference between the error sensitivity in the weak- and strong-impostor conditions. As shown in Figure

S4, our simulations indicate that we were able to detect an interaction of 8 percentage points in confidence with 80% power.

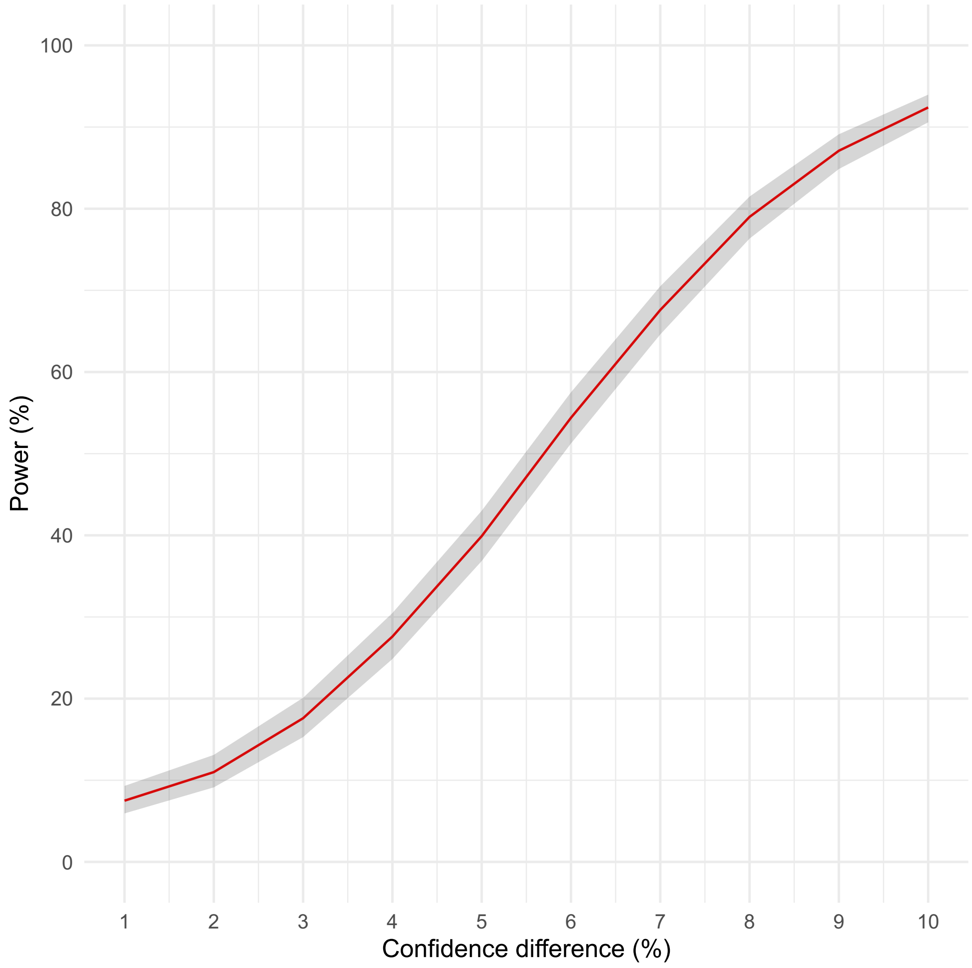

Figure S4. Sensitivity power curve illustrating the achieved power as a function of effect size for n=200 on confidence in Experiment 3, based on 1000 simulations per effect size. Confidence difference denotes the magnitude of the interaction between impostor strength and incorrect anomaly responses in confidence percentage points.

B. Instructions

Experiment 1-2

Please read these instructions carefully!

In this experiment you will have to answer 20 multiple-choice trivia questions and 2 practice questions.

For every multiple choice question you will be presented with four answer options but **you can only pick one answer. Please respond as accurately as you can.**

Some of the questions are impossible to answer. In that case, select the answer option: 'This question can't be answered in this form.'

If you don't know the answer to a question, select the response option 'Don't know'.

To clarify the difference between 'Don't know' and 'This question can't be answered in this form.', take a look at the example questions below:

What is the name of former president's Obama's oldest son?

Charles
Jonathan
This question can't be answered in this form.
Don't know.

The above question cannot be answered because Obama doesn't have a son; he only has two daughters. So, the correct answer option to this question is: 'This question can't be answered in this form.'

Here is a different example:

What is the name of former president's Obama's oldest daughter?

Sasha
Malia
This question can't be answered in this form.
Don't know.

In the above example, the question can be answered, since Obama does have an oldest daughter. The correct answer option is 'Malia'. However, if you do not know the answer to this question, you should select 'Don't know'.

> Critically, in this study we want to know what your **initial, intuitive response** to the questions is and **how you respond after you have thought about these questions for some more time.**
>
> First, we want you to respond with the **very first answer that comes to mind.** You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible.
>
> To make sure that you answer as fast as possible, a time limit was set for the first response, which is going to be **5 seconds** (Experiment 1)/**4 seconds** (Experiment 2). When there is 1 second left, the background colour will turn to yellow to let you know that the deadline is approaching. Please make sure to **answer before the deadline passes.**
>
> Next, **the question will be presented again** and you can take all the time you want to actively reflect on it. Once you have made up your mind you give your **final response.**
>
> After you made your choice and clicked on it, you will be automatically taken to the next page.
>
> After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.
>
> We are going to clarify all of this with a couple of practice questions.

> First, a fixation cross will appear. Then, the question and the four answer options will appear. You then enter your first hunch as fast as possible before the deadline. Next, the question will be presented again and you can take all the time to reflect on it and enter your final response.
>
> After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.

Participants were then given two practice trials, without the concurrent load task. They were then introduced to the load task.

> You will also need to **memorize a pattern** while you respond to the trivia questions.
>
> You will see a grid with crosses and you will have to memorize their location.
>
> You will first practice with 2 patterns without a trivia question.
>
> The pattern will be displayed for **2 seconds** and then you will have to select it among **4 different patterns.**

Participants were then given two practice trials for the cognitive load task, without the multiple-choice questions. They were then presented with the following instructions:

> In the actual study you will need to memorize the pattern while you respond to the trivia question. The pattern is briefly presented before each question.

> The difficulty of the pattern might vary. Always try to memorize as many crosses as possible. Each cross counts!
>
> We know that it is not always easy to memorize the pattern while you are also thinking about the trivia question. The most important thing is to correctly memorize the pattern.
>
> First, **try to concentrate on the memorization task**, and then try to answer the question accurately.
>
> As a next step, you can practice this with two questions.

After those two last practice trials, participants were presented with the following instructions:

> Ok, this is the end of practice!
>
> During the experiment, the questions will be presented to you one after the other and you should not pause between them. After the first 10 questions, you can take a short break.
>
> Remember, some of the questions are impossible to answer. In that case, select the answer option: 'This question can't be answered in this form.' If you don't know the answer to a question, select the response option 'Don't know'.

Experiment 3

> **Please read these instructions carefully!**
>
> In this experiment, you will have to respond to 20 multiple-choice trivia questions and 8 practice questions.
>
> We want to know what your **intuitive response** to these questions is and how you respond if you are **taking more time to think** about the questions.
>
> We will always instruct you as to whether we want you to answer the question **intuitively** or whether you can **take your time to deliberate on your decision**.

Then the trivia task was introduced as in the previous experiments, using the same example (see above). Participants then received the following instructions:

> In the **intuitive** trials, we are interested in your **initial, intuitive response**. We want you to respond with the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible.
>
> To make sure that you answer as fast as possible, **a time limit was set for the intuitive trials response**, which is going to be **5 seconds**. When there is 1 second left, the background color will turn to **yellow** to let you know that the deadline is approaching. Please make sure to **answer before the deadline passes**.

> For the **deliberative trials**, you can **take the time to actively reflect on the question** before submitting your response. Here we want you to think as deeply as possible before you give your answer.

> To better understand the experiment, you are first going to solve some practice questions.
>
> First, a fixation cross will appear. Then, the question and the four answer options will appear.
>
> After you made your choice and clicked on it, you will be automatically taken to the next page.
>
> In both **intuitive** and **deliberative** trials, we will also ask you to indicate your **confidence** in the correctness of your response.

> In the next part, you will have to give the **first answer that comes to mind, as quickly as possible**.

Participants then answered the first practice question without the cognitive load:

> That was the first practice question.
>
> We will present you with a second practice question to make sure that the procedure is really clear.
>
> Keep in mind that it is really important to answer the question as fast as possible as you are providing your **intuitive response**.

After they answered the second practice question, the following instructions were displayed to introduce the load task:

> In the **intuitive trials**, you will also need to **memorize a pattern** while you respond to the trivia questions.
>
> You will see a grid with crosses, and you will have to memorize their location.
>
> You will first practice with 2 patterns without a trivia question.
>
> The pattern will be displayed for **2 seconds**, and then you will have to select it among **4 different patterns**.

They then practiced the load task without answering the trivia questions.

> In the **intuitive trials** of the actual study, you will need to memorize the pattern while you respond to the trivia question. The pattern is briefly presented before each question.

> The difficulty of the pattern might vary. Always try to memorize as many crosses as possible. Each cross counts!
>
> We know that it is not always easy to memorize the pattern while you are also thinking about the trivia question. The most important thing is to correctly memorize the pattern.
>
> First, **try to concentrate on the memorization task**, and then try to answer the question accurately.
>
> As a next step, you can practice this with two trivia questions.

After that, they had to complete two intuitive block trials with both the deadline and the load task. Then they saw the following instructions, which introduced the deliberative block:

> In the next part, you can take as much time as you need to reflect on the trivia question to give your **deliberative answer**.
>
> Here you won't have to memorize any pattern while you answer the trivia question.
>
> We'll let you practice with two trivia questions.

After that, they had to answer the first deliberative practice questions before receiving the following instructions:

> That was the first practice question.
>
> We will present you with a second practice question to make sure that the procedure is really clear.
>
> Keep in mind that you can take as much time as you need to reflect on the question to give your **deliberative answer**.

After answering the second deliberative block questions, they received the following instructions:

> This is the end of the practice session. You are now ready to start the experiment.
>
> You will be presented with 2 blocks, in which you will need to answer 10 trivia questions each time.
>
> In the **intuitive thinking block** you will need to give your answer **intuitively, as fast as possible**. In addition, you'll have to memorize a pattern while you answer the trivia questions.
>
> In the **deliberative thinking block**, you can take **as much time as you need to reflect on the question** before giving your answer. Here you won't have to remember any pattern while you answer the trivia questions.

Please make sure to stay **maximally focused** throughout the study.

Depending on the counterbalancing order, they then saw one of these two prompts:

- You will now start a block of 10 intuitive thinking trials. Give the first answer that comes to mind, as quickly as possible, while memorizing the pattern.
- You will now start a block of 10 deliberative thinking trials. You can take all the time you want to reflect on the trivia question before giving your response.

Experiment 3 Norming Study

Please read these instructions carefully!

In this experiment, you will be presented with 25 items. Each item will include an undistorted, correct trivia question (along with the correct answer between brackets), and two distorted versions of this question. In the distorted versions, one or more words of the original question have been replaced by one or more "impostor" words. We know that people often fail to notice such replacements when the impostor word is very similar to the undistorted word.

Your task is to indicate **how similar each distorted sentence is to the original undistorted question on a scale ranging from 0 (Not at all similar) to 100 (Extremely similar).**

To clarify, take a look at the following example:

The undistorted question is: "What is the name of Harry Potter's female best friend, in the famous fantasy NOVEL by J.K. Rowling? (answer: Hermione Granger)".

How similar is each distorted sentence to the original undistorted question?

Please type a number from 0 (Not at all similar) to 100 (Extremely similar) for each sentence.

What is the name of Harry Potter's female best friend, in the famous fantasy POEM by J.K. Rowling?

What is the name of Harry Potter's female best friend, in the famous fantasy SONATA by J.K. Rowling?

In this example, you might think that the first version ("POEM") is more similar to the original version than the second one ("SONATA"), and that here "POEM" can go unnoticed more easily. In this case, you would have to give a higher similarity rating to the first distorted sentence than to the second one.

C. Estimated models without trial exclusion

In this section, we report the results of the analyses where we do not exclude trials with a missed cognitive load and recode trials with a missed deadline as incorrect. For confidence analyses, note that we still had to exclude trials with a missed deadline (but not with a missed cognitive load) since participants did not provide their confidence on these trials and we do not have a default value to replace them with as we had for accuracy (i.e., incorrect).

Overall, there were few differences in significance in these additional analyses. However, note that the interaction between illusion strength and anomaly-correct responses on confidence was significant in the revised illusion strength analysis of Experiment 2 (*p* = .04; Table S8). Additionally, in the analysis of confidence as a function of direction of change in Experiment 1, the difference between the reference no-anomaly "11" category and the (very rare) anomaly "10" trials was no longer significant (*p* = .08; see Table S3).

Table S1

Binomial generalized mixed-effects model on accuracy as a function of anomaly presence and response stage in Experiment 1, using sum coding

accuracy ~ 1 + anomaly + response stage + anomaly:response stage +
(1 + anomaly | subject) + (1 + anomaly | item)

Fixed effects

Predictors	df	χ^2	p.value	Cohen's d
Anomaly	1	42.52	**< .001**	-1.17
Response stage	1	48.17	**< .001**	0.19
Anomaly:Response stage	1	18.44	**< .001**	0.12

Random effects

Group	Parameter	SD
Subject	(Intercept)	1.03
Item	(Intercept)	0.72
Subject	Anomaly	0.81
Item	Anomaly	0.73
Subject	Cor (Intercept x Anomaly)	0.38
Item	Cor (Intercept x Anomaly)	-0.49

Model fit

Metric	R^2 (Marginal)	R^2 (Conditional)
	0.49	0.66

N	Subject	Item	Observations
	100	20	4,000

Note. p-values were obtained through bootstrapping on the comparison between a full model and a reduced model without the variable of interest, using type III sum of squares.

Table S2.

Mixed-effects beta-regression on initial confidence ratings as a function of anomaly and accuracy in Experiment 1, using dummy coding

confidence ~ 1 + anomaly + accuracy + anomaly:accuracy + (1 | item)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly correct)	3.42	[3.07, 3.81]	**< .001**	0.68
Anomaly incorrect	0.64	[0.56, 0.72]	**< .001**	-0.25
Anomaly correct	1.14	[0.95, 1.37]	.12	0.07
No-anomaly incorrect	0.20	[0.16, 0.26]	**< .001**	-0.88

Random effects		
Group	**Parameter**	**SD**
Item	(Intercept)	0.18

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.43	0.51

N	**Subject**	**Item**	**Observations**
	100	20	1,865

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

Table S3

Beta-regression results contrasting the initial confidence ratings for "11" control no-anomaly trials with anomaly trials for each direction of change category in Experiment 1, using dummy coding

confidence ~ 1 + response category + (1 | item)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly "11")	3.95	[3.51, 4.45]	**< .001**	0.76
Anomaly "11"	1.15	[0.94, 1.39]	.15	0.08
Anomaly "00"	0.77	[0.68, 0.87]	**< .001**	-0.15
Anomaly "01"	0.20	[0.16, 0.25]	**< .001**	-0.89
Anomaly "10"	0.51	[0.23, 1.06]	.08	-0.37

Random effects		
Group	**Parameter**	**SD**
Item	(Intercept)	0.16

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.55	0.62

N	**Subject**	**Item**	**Observations**
	100	20	1,738

Note. p-values and confidence intervals were obtained using parametric bootstrap. OR = odds ratio. An OR superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence relative to the to the "11" no-anomaly control condition.

To further explore the role of response change on initial confidence, we compared trials in which participants changed their response between the initial and final stages ("01" and "10" categories) to those in which they did not change their response ("00" and "11" categories). Replicating previous findings, a post-hoc contrast showed that initial confidence was significantly lower in change trials than in no-change trials: *OR* = 2.95, 95% CI [1.99, 4.44], *p* < .001, Cohen's *d* = 0.63.

Table S4

Beta-regression predicting initial confidence ratings in Experiment 1 as a function of standardized illusion strength, response group, and their interaction, using dummy coding

confidence ~ 1 + group + illusion strength + illusion strength:group

Predictors	OR	95% CI	Z value	p.value	Cohen's d
Intercept (no-anomaly correct)	3.58	[3.49, 3.68]	26.4	**<.001**	0.7
Anomaly correct	1.09	[1, 1.18]	1.83	.07	0.05
Anomaly incorrect	0.61	[0.51, 0.71]	-9.51	**<.001**	-0.27
Illusion strength	1.14	[1.1, 1.17]	7.08	**<.001**	0.07
Anomaly correct:Illusion strength	0.85	[0.76, 0.95]	-3.36	**<.001**	-0.09
Anomaly incorrect:Illusion strength	1.17	[1.08, 1.25]	3.45	**<.001**	0.08

Model fit

Metric	Pseudo R^2
	0.12

N	Subject	Item	Observations
	100	20	1,760

Note. p-values were obtained using Wald Z-tests with clustered standard errors by participants. OR = odds ratio. An OR superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

Table S5

Binomial generalized mixed-effects model on accuracy as a function of anomaly presence and response stage in Experiment 2, using sum coding

accuracy ~ 1 + anomaly + response stage + anomaly:response stage + (1 + anomaly | subject) + (1 + anomaly || item)

Fixed effects				
Predictors	**df**	**χ^2**	**p.value**	**Cohen's d**
Anomaly	1	43.89	**< .001**	-1.03
Response stage	1	104.90	**< .001**	0.26
Anomaly:Response stage	1	35.48	**< .001**	0.15

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.94
Item	(Intercept)	0.70
Subject	Anomaly	0.70
Item	Anomaly	0.61
Subject	Cor (Intercept x Anomaly)	0.69

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.45	0.61

N	Subject	Item	Observations
	100	20	4,000

Note. p-values were obtained through bootstrapping on the comparison between a full model and a reduced model without the variable of interest, using type III sum of squares.

Table S6

Mixed-effects beta-regression on initial confidence ratings as a function of anomaly and accuracy in Experiment 2, using dummy coding

confidence ~ 1 + anomaly + accuracy + anomaly:accuracy + (1 | subject) + (1 | item)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly correct)	2.94	[2.62, 3.31]	**< .001**	0.59
Anomaly incorrect	0.62	[0.54, 0.70]	**< .001**	-0.26
Anomaly correct	1.12	[0.88, 1.42]	.33	0.06
No-anomaly incorrect	0.29	[0.23, 0.37]	**< .001**	-0.68

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.24
Item	(Intercept)	0.12

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.31	0.48

N	**Subject**	**Item**	**Observations**
	100	20	1,703

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

Table S7

Beta-regression results contrasting the initial confidence ratings for “11” control no-anomaly trials with anomaly trials for each direction of change category in Experiment 2, using dummy coding

confidence ~ 1 + response category + (1 | subject) + (1 | item)

Fixed effects

Predictors	OR	95% CI	p.value	Cohen’s d
Intercept (no-anomaly "11")	3.39	[3.03, 3.81]	< **.001**	0.67
Anomaly "11"	1.18	[0.93, 1.50]	.16	0.09
Anomaly "00"	0.75	[0.65, 0.86]	< **.001**	-0.16
Anomaly "01"	0.23	[0.18, 0.28]	< **.001**	-0.82
Anomaly "10"	0.34	[0.16, 0.69]	< **.001**	-0.59

Random effects

Group	Parameter	SD
Subject	(Intercept)	0.21
Item	(Intercep)	0.11

Model fit

Metric	R^2 (Marginal)	R^2 (Conditional)
	0.49	0.59

N	Subject	Item	Observations
	100	20	1,539

Note. p-values and confidence intervals were obtained using parametric bootstrap. OR = odds ratio. An OR superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence relative to the to the “11” no-anomaly control condition.

To further explore the role of response change on initial confidence, we compared trials in which participants changed their response between the initial and final stages (“01” and “10” categories) to those in which they did not change their response (“00” and “11” categories). Replicating previous findings, a post-hoc contrast showed that initial confidence was significantly lower in change trials than in no-change trials: *OR* = 3.38, 95% CI [2.28, 5.12], *p* < .001, Cohen’s *d* = 0.69.

Table S8

Mixed-effects beta-regression predicting initial confidence ratings in Experiment 2 as a function of standardized illusion strength, response group, and their interaction, using dummy coding

confidence ~ 1 + group + illusion strength + illusion strength:group + (1 | subject)

Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly correct)	2.99	[2.67, 3.38]	**< .001**	0.60
Anomaly correct	1.10	[0.86, 1.42]	.40	0.05
Anomaly incorrect	0.61	[0.53, 0.70]	**< .001**	-0.27
Illusion strength	1.17	[1.06, 1.29]	**< .001**	0.09
Anomaly correct:Illusion strength	0.79	[0.63, 0.99]	**.04**	-0.13
Anomaly incorrect:Illusion strength	1.06	[0.94, 1.22]	.36	0.03

	Random effects	
Group	**Parameter**	**SD**
Subject	(Intercept)	0.22

	Model fit	
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.28	0.42

N	**Subject**	**Item**	**Observations**
	100	20	1,567

Note. p-values and confidence intervals were obtained using parametric bootstrap. OR = odds ratio. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

Table S9

Binomial generalized mixed-effects model on accuracy as a function of impostor strength, anomaly presence and response block in Experiment 3, using sum coding

accuracy ~ 1 + impostor + response block + anomaly + impostor:response block + impostor:anomaly + response block:anomaly + impostor:response block:anomaly + (1 + anomaly | subject) + (1 + anomaly | item)

Fixed effects				
Predictors	**df**	**χ^2**	**p.value**	**Cohen's d**
Impostor	1	9.43	**.003**	-0.16
Response Block	1	116.08	**< .001**	-0.28
Anomaly	1	44.51	**< .001**	0.97
Impostor:Response Block	1	0.10	.79	-0.00
Impostor:Anomaly	1	20.14	**< .001**	0.15
Response Block:Anomaly	1	19.89	**< .001**	0.12
Impostor:Response Block:Anomaly	1	0.01	.93	0.01

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	1.14
Item	(Intercept)	0.51
Subject	Anomaly	0.48
Item	Anomaly	0.58
Subject	Cor (Intercept x Anomaly)	-0.55
Item	Cor (Intercept x Anomaly)	0.15

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.43	0.61

N	**Subject**	**Item**	**Observations**
	200	20	4,000

Note. p-values were obtained through bootstrapping on the comparison between a full model and a reduced model without the variable of interest, using type III sum of squares.

Table S10

Mixed-effects beta-regression on confidence as a function of response group and impostor strength in Experiment 3, using dummy coding

confidence ~ 1 + impostor + group + impostor:group + (1 | subject) + (1 | item)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly correct impostor strong)	3.96	[3.32, 4.55]	**< .001**	0.76
Impostor weak	1.00	[0.82, 1.23]	.98	0.00
Anomaly correct	0.99	[0.73, 1.29]	.94	-0.01
Anomaly incorrect	0.62	[0.53, 0.73]	**< .001**	-0.26
Anomaly correct:Impostor weak	1.09	[0.73, 1.57]	.71	0.05
Anomaly incorrect:Impostor weak	0.91	[0.72, 1.13]	.40	-0.05

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.48
Item	(Intercept)	0.14

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.15	0.71

N	**Subject**	**Item**	**Observations**
	200	20	1,710

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

Table S11

Mixed-effects beta-regression on confidence as a function of response group and standardized illusion strength in Experiment 3, using dummy coding

confidence ~ 1 + group + illusion strength + illusion strength:group + (1 | subject)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (no-anomaly correct)	4.02	[3.62, 4.48]	**< .001**	0.77
Anomaly correct	0.98	[0.81, 1.20]	.88	-0.01
Anomaly incorrect	0.58	[0.51, 0.65]	**< .001**	-0.30
Illusion strength	1.11	[1.02, 1.22]	**.01**	0.06
Anomaly correct:Illusion strength	0.87	[0.72, 1.05]	.21	-0.08
Anomaly incorrect:Illusion strength	1.15	[1.02, 1.31]	**.017**	0.08

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.48

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.21	0.72

N	**Subject**	**Item**	**Observations**
	200	40	1,710

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in confidence.

D. Additional Models

Table S12

Log-linear mixed model contrasting the latencies of the one-response pre-test and the initial response stage of Experiment 1 on correct anomaly trials, using dummy coding

log(RT) ~ 1 + experiment + (1 | subject) + (1 | item)

Fixed effects					
Predictors	**exp(Est)**	**SE**	**t val.**	**df**	**p.value**
Intercept (one-response)	6.35	0.05	39.65	61.74	**< .001**
Experiment 1	0.59	0.05	-9.48	96.31	**< .001**

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.21
Item	(Intercept)	0.11
Residuals		0.30

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.32	0.58

N	**Subject**	**Item**	**Observations**
	107	20	363

Note. p-values were obtained using Kenward-Roger d.f.

Table S13

Binomial generalized mixed-effects model contrasting the accuracy of the one-response pre-test and the final response stage of Experiment 1, using dummy coding

accuracy ~ 1 + experiment + (1 | subject) + (1 | item)

Fixed effects				
Predictors	**OR**	**95% CI**	**p.value**	**Cohen's d**
Intercept (one-response)	1.93	[1.41, 2.61]	**< .001**	0.36
Experiment 1	1.06	[0.79, 1.43]	.73	0.03

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.65
Item	(Intercept)	0.44

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.00	0.16

N	**Subject**	**Item**	**Observations**
	150	20	2,636

Note. p-values and confidence intervals were obtained using parametric bootstrap. *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in the odds of giving the correct response.

Table S14

Log-linear mixed model contrasting the latencies of Experiment 1 and Experiment 2 in the initial response stage, using dummy coding

log(RT) ~ 1 + experiment + (1 | subject) + (1 | item)

Fixed effects					
Predictors	**exp(Est)**	**SE**	**t val.**	**df**	**p.value**
Intercept (Experiment 1)	3.12	0.02	53.83	55.61	**< .001**
Experiment 2	0.92	0.02	-4.19	197.94	**< .001**

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.14
Item	(Intercept)	0.07
Residuals		0.20

Model fit		
Metric	**R² (Marginal)**	**R² (Conditional)**
	0.03	0.39

N	**Subject**	**Item**	**Observations**
	200	20	3,005

Note. *p*-values were obtained using Kenward-Roger d.f.

Table S15

Log-linear mixed model contrasting the latencies of the intuitive and the deliberative response blocks in Experiment 3 on correct anomaly trials, using dummy coding

log(RT) ~ 1 + response block + (1 | subject) + (1 | item)

Fixed effects					
Predictors	**exp(Est)**	**SE**	**t val.**	**df**	**p.value**
Intercept (intuitive block)	3.87	0.05	27.62	67.06	**< .001**
Deliberative block	1.72	0.04	15.06	474.32	**< .001**

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.26
Item	(Intercept)	0.14
Residuals		0.34

Model fit		
Metric	**R^2 (Marginal)**	**R^2 (Conditional)**
	0.23	0.56

N	**Subject**	**Item**	**Observations**
	158	20	546

Note. *p*-values were obtained using Kenward-Roger d.f.

E. Statistical Analysis Procedure

Random structure specification

To find the best random structure for each analysis, we identified the maximal model supported by the data (i.e., capable of converging), before performing backward stepwise elimination using the likelihood ratio test. This allowed us to find the optimal model without sacrificing statistical power (Matuschek et al., 2017). When the random structure was unsupported or did not improve model fit, we used clustered standard errors to account for the non-independence of our data (Cameron & Miller, 2015).

Significance evaluation and contrast coding

For generalized mixed models, we used parametric bootstrapping with 1000 iterations per model to evaluate the significance of the fixed effects (Booth, 1995), which does not rely on any traditional statistical distribution (e.g., χ^2 distribution). In the linear mixed models on reaction times, we used the Kenward-Roger's F test (Kenward & Roger, 1997) to assess the significance of the fixed effects. In the case of beta regression models with clustered standard errors (i.e., when the random structure was unsupported), we used Wald Z-tests to assess significance.

For each analysis, we specify the contrast coding scheme we used based on the analysis requirements (Brehm & Alday, 2022). When comparing different variable levels to a specific group (e.g., the control no-anomaly group), we applied treatment (i.e., dummy) coding to sidestep the need for post-hoc tests (Schad et al., 2020). Alternatively, when this comparison was not necessary, we used sum coding in conjunction with Type III sum of squares to derive inferences similar to traditional ANOVA models. In this case, bootstrapping was used on the comparison between a full model and a reduced model without the variable of interest (Halekoh & Højsgaard, 2014), to compute a reference distribution of χ^2 values[1].

In the event of a significant interaction, we conducted post-hoc tests on the estimated marginal means to explore it further, using the Holm-Bonferroni method to correct for multiple comparisons. For generalized (mixed) models, we report effect sizes by converting the odds ratio to Cohen's *d* (Borenstein et al., 2009), reporting both the point estimates and their 95% confidence intervals.

[1] We do not report confidence intervals for the effect sizes of these analyses since they do not provide confidence intervals for the regression coefficients.

F. Distribution of individual non-correction rates and initial errors sensitivity measures

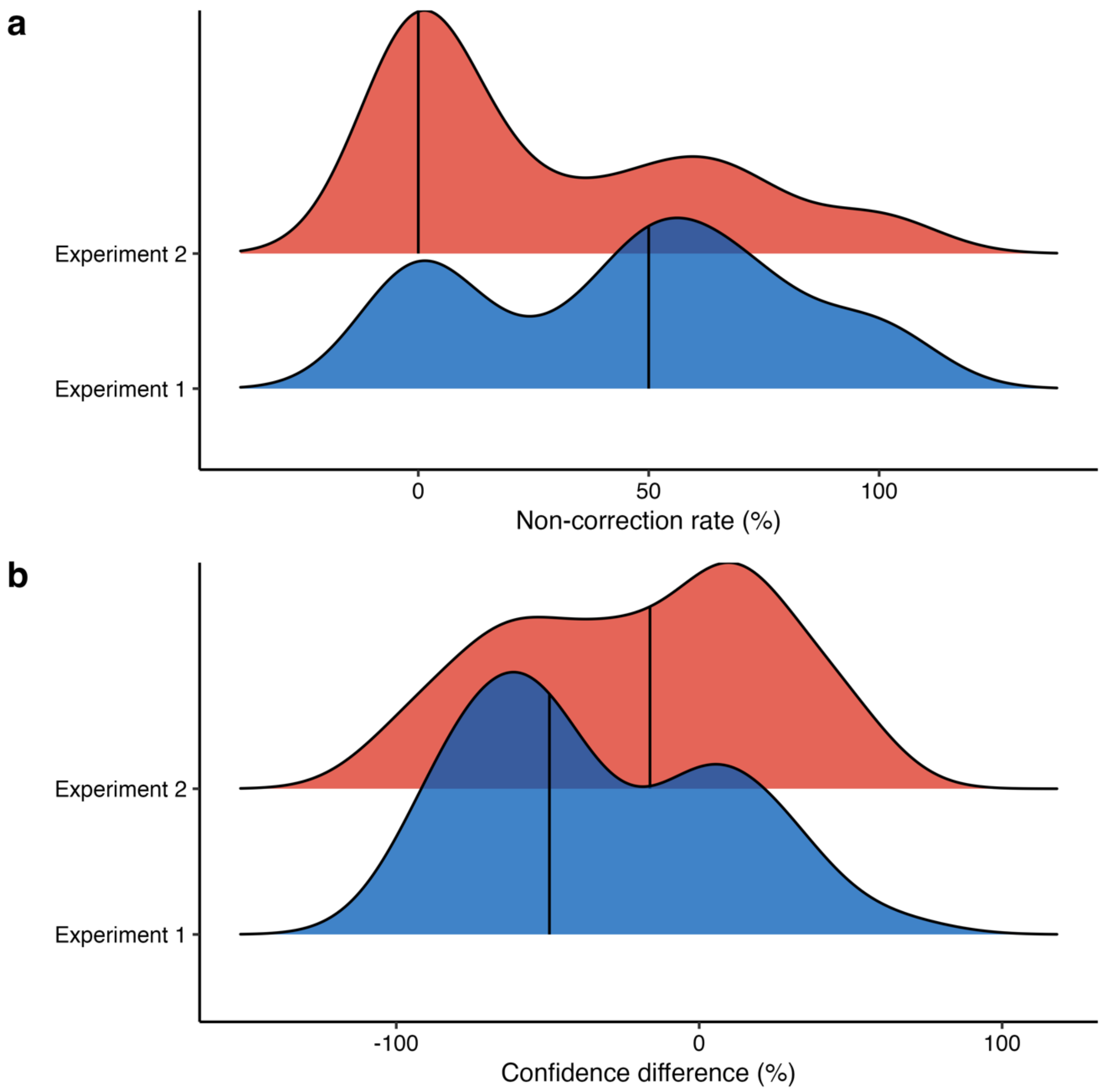

Figure S5. Ridgeline density plots of non-correction rates and initial error sensitivity measures for the two experiments. **a**) Individual non-correction rates for anomaly problems. **b**) Initial confidence difference between the correct control no-anomaly trials and the incorrect anomaly trials. Black lines indicate the median.

Table S16

Binomial generalized mixed-effects model on the probability of giving a "11" response (vs. a "01" response) in anomaly trials contrasting Experiment 1 and Experiment 2, using dummy coding

correct int ~ 1 + experiment + (1 | subject) + (1 | item)

Fixed effects			
Predictors	**OR**	**95% CI**	**p.value**
Intercept (Experiment 1)	0.97	[0.60, 1.53]	.90
Experiment 2	0.41	[0.25, 0.65]	**< .001**

Random effects		
Group	**Parameter**	**SD**
Subject	(Intercept)	0.66
Item	(Intercept)	0.85

Model fit			
Metric		**Pseudo-R² (Marginal)**	**Pseudo-R² (Total)**
		0.04	0.29
N	**Subject**	**Item**	**Observations**
	154	20	525

Note. p-values and confidence intervals were obtained using parametric bootstrap. This model is based only on trials with a final correct response (i.e., "11" or "01" responses). *OR* = odds ratio. An *OR* superior/inferior to 1 indicates the magnitude of the increase/decrease in the odds of giving a "11" response vs. a "01" response.