Intuitive Insight: Fast Associative Processes Drive Sound Creative Thinking

Jérémie Beucler[a], Wim De Neys[a]

[a] Université Paris-Cité, LaPsyDÉ, CNRS, F-75005 Paris, France

Email addresses: J.B.: jeremie.beucler@gmail.com; W.D.N.: wim.de-neys@parisdescartes.fr

Correspondence concerning this article should be addressed to Jérémie Beucler , Sorbonne, LaPsyDÉ, 46 rue Saint-Jacques, 75005 Paris, France. Email: jeremie.beucler@gmail.com

Preprint Notice

This manuscript has been accepted for publication in Cognition. The final version may slightly differ from this preprint.

Abstract

Convergent thinking, the ability to find a single optimal solution to a well-defined problem, is considered a core component of creativity, and is often assumed to rely on controlled, deliberative processes. We tested this assumption using the Compound Remote Associates (CRA) test, where participants have to find a word that connects three seemingly unrelated words (e.g., "river, note, account"; solution: "bank"). We implemented a two-response paradigm wherein participants provided an initial, intuitive response (under cognitive load and time constraints to minimize deliberation), followed by a final, deliberate response. Our findings reveal that, in most cases, extended deliberation was not necessary for sound thinking—correct final responses were typically preceded by accurate intuitive responses produced under time pressure and cognitive load. By using large language models and semantic network modeling, we found that items with a smaller semantic search space are better solved intuitively, and that participants with a more efficient and flexible semantic memory structure display higher intuitive performance on the CRA. These results suggest that effective problem-solving in creative tasks may often rely on fast, automatic associative processes within semantic memory, without necessarily requiring extended deliberation.

Keywords: dual-process; creativity; convergent thinking; associative thinking

1. Introduction

Does sound thinking require deliberation, or can it emerge intuitively? Dual-process theories have been highly influential in cognitive psychology, particularly in explaining biases and errors in reasoning. According to popular dual-process accounts, human cognition results from a dynamic interplay between fast and effortless, intuitive ("Type 1") processing and slower, more effortful, deliberate ("Type 2") processing (Evans & Stanovich, 2013; Kahneman, 2011; Stanovich & West, 2000). Although intuitive processes often cue valid responses, they rely on heuristic shortcuts that can sometimes lead us astray. In contrast, deliberate processes allow individuals to detect and correct erroneous intuitive responses. Traditionally, this view implies that sound thinking essentially involves engaging deliberation to override initial misleading intuitions. Beyond reasoning, dual-process frameworks have been applied to a wide range of cognitive domains, including semantic illusions (Koriat, 2017; Mata et al., 2013), moral judgment (Bago & De Neys, 2019a; Greene & Haidt, 2002), prosocial behavior (Rand et al., 2012), and notably, creativity (Allen & Thomas, 2011; Barr et al., 2014; Cassotti et al., 2016; Sowden et al., 2019).

However, recent dual-process studies in the reasoning field question the classic dual-process conceptualization. In these studies, participants typically have to tackle logico-mathematical "bias" problems, such as the well-known bat-and-ball task: "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"). The classic dual-process assumption is that correct responding requires a deliberate evaluative phase in which initial, intuitively cued solutions (e.g., "10 cents") are corrected. To disentangle the distinct roles of intuitive and deliberative processes, researchers have used a two-response paradigm (Thompson et al., 2011). In the first, intuitive stage of this paradigm, participants are required to provide the first answer that comes to their mind as quickly as possible. Immediately after, they have all the

time they want to deliberate and pick their final response. To ensure that the initial response does not rely on deliberation, participants have to answer under time pressure while performing a concurrent cognitive load task hampering their cognitive resources (Bago & De Neys, 2017). Given that deliberation is assumed to require time and cognitive resources, these constraints minimize the possibility that people engage in it.

Contrary to the predictions of the classic dual-process framework, results indicate that sound reasoners are often able to produce correct responses in the initial, intuitive response stage where deliberation is minimized (Bago & De Neys, 2017, 2019b; Burič & Konrádová, 2021; Burič & Šrol, 2020; Raoelison et al., 2020; Thompson & Johnson, 2014). It has been argued that this suggests that deliberate control processes might be less critical for sound thinking than generally assumed (De Neys, 2023). Hence, sound reasoners may be primarily characterized by their ability to intuit correctly rather than by their capacity to deliberately correct erroneous intuitions (Raoelison et al., 2020; Reyna, 2012; Thompson et al., 2018).

Since the dual-process framework is often posited as a general model of human cognition (Evans & Stanovich, 2013; Kahneman, 2011), we believe it is important to assess the generalizability of these findings in other cognitive domains (e.g., Beucler et al., 2025; Voudouri et al., 2023). Creativity provides an ideal domain for examining this question. Much like revised reasoning-based dual-process models, recent creativity models appear more open to the possibility of a greater role for intuition in the creative process. Evidence for such a role comes from studies showing that people can often arrive at solutions intuitively, or at least recognize when problems are solvable, without engaging in deliberate analysis. For instance, Topolinski and Strack (2009) found that participants were able to distinguish solvable from unsolvable word triads based on intuitive impressions. Similarly, research on insight problem-solving, when a solution emerges

suddenly with an "Aha!" experience, has shown that certain creativity problems can be solved intuitively (e.g., Stuyck et al., 2022). Thus, recent theories in creativity research emphasize a dynamic interplay between intuitive and deliberate processes (Benedek et al., 2023; Volle, 2018), whereby bottom-up associative and top-down controlled processes continuously interact to produce a creative solution.

To directly assess the extent of the role of deliberation in creative thinking, we adapted the two-response paradigm to the Compound Remote Associates test (CRA; Bowden & Jung-Beeman, 2003). In this classic "convergent-thinking" task, participants are presented with three problem words (e.g., river, note, account) and must find a fourth word that can be appended to each of the three words to create three new meaningful compound words (e.g., *river*bank, *bank*note, *bank* account). Participants were required to provide their initial responses as quickly as possible while under time pressure (8 seconds in Experiment 1; 6 seconds in Experiment 2) and concurrent cognitive load (memorization of a 3x3 matrix in Experiment 1 and a 4x4 matrix in Experiment 2; see Bago & De Neys, 2017, 2019a for validation). Subsequently, they were allowed to deliberate and give their final responses. To ensure our items matched the difficulty level of previous studies, we selected CRA items from Bowden and Jung-Beeman (2003) and conducted a pilot study to confirm item difficulty.

Critically, our objective was twofold. First, we aimed to measure the extent of sound intuitive thinking in the CRA. Here, sound intuitive thinking is defined in a strictly operational sense as the production of a correct response at the initial response stage of the two-response paradigm, under time pressure and concurrent cognitive load, regardless of the specific cognitive processes involved. Second, we sought to explain how such correct intuitions arise by investigating the underlying associative processes that may support them. Indeed, creative thinking is often assumed

to hinge on associative mechanisms that link remote concepts to generate novel ideas (Beaty & Kenett, 2023; Kenett et al., 2014; Mednick, 1962; Rossmann & Fink, 2010). Such associative dynamics are typically linked to intuitive, "Type 1" processes within the traditional dual-process view of reasoning (Evans & Stanovich, 2013; Kahneman, 2011), though this simple mapping between intuitive and associative processes is likely an oversimplification in the context of creativity (Ovando-Tellez et al., 2024; Sowden et al., 2019). A key goal of the present work is to empirically refine our understanding of the relationship between associative processes and intuitive thinking.

Importantly, both item-level and individual-level factors may influence the ease with which intuitive responses emerge. At the item level, prior work has shown the semantic similarity between cue words and the solution word partly predicts performance in convergent-thinking tasks (Marko et al., 2019). Semantic similarity refers to the proximity in meaning between two words or groups of words. For example, in the CRA item "extinguisher, truck, camp", the solution "fire" is semantically close to the cues, whereas in "baby, spring, cap", the solution "shower" is more distant, making the item harder to solve.

At the individual level, creative performance has been linked to differences in how knowledge is organized and accessed in semantic memory (e.g., Kenett et al., 2014, 2018; Luchini et al., 2023). Semantic memory is often modeled as a network of interconnected concepts, in which related concepts are closer and more densely interconnected than distant ones. Individual differences in the structure of these semantic networks, such as higher connectivity or shorter paths between concepts, are thought to underly greater creative performance. For some individuals, concepts that are typically considered remote may thus be more closely interconnected in their semantic memory networks, thereby reducing the reliance on deliberate processes. Complementing

this network-based view of semantic memory, recent work has emphasized the importance of dynamic aspects of associative search. In particular, the forward flow measure was developed to capture how efficiently individuals explore semantic space during free association (Gray et al., 2019). A higher forward flow reflects greater movement through semantic space and has been shown to predict creative performance (Beaty et al., 2021; Gray et al., 2019).

Together, these perspectives suggest that both problem structure and the organization and dynamics of semantic memory determine whether solutions are reached intuitively or require deliberation. To test this, we adopted recent advances in computational modeling of creativity (Beaty & Kenett, 2023). At the item level, we examined whether semantic similarity between the three cue words and the solution word of the problem (via word embeddings) predicted intuitive versus deliberative solutions for each item. At the participant level, we investigated whether individual differences in semantic memory structure and associative search dynamics help explain why some individuals are more successful at intuitive problem solving. To this end, we combined network modeling of a verbal fluency task to characterize the structure of semantic memory, and the forward flow measure to index the dynamics of associative exploration during free association.

We conducted two experiments: Experiment 1 introduced the two-response paradigm to identify the role of intuitive and deliberate processing in the CRA task. Experiment 2 validated the results (with a tighter deadline and a more demanding cognitive load task to prevent any possible residual deliberation, see Bago & De Neys, 2019a) and introduced the additional verbal fluency task to model interindividual differences in semantic memory structure and exploration.

2. Methods

2.1. Transparency and Openness

The research question and study design were preregistered on the OSF platform (Experiment 1: https://osf.io/32m7f/?view_only=3c119f6744f8442ba3fad4e324035f7b; Experiment 2: https://osf.io/6f47j/?view_only=a6bc40550fe04f6cb4220cd8d4f0141a). No specific analyses were preregistered. All data, material, and analysis scripts can be retrieved from: https://osf.io/gfsb8/?view_only=81b5433fef4040dc949332ed66aeb3c0.

2.2. Participants

Participants were native English speakers recruited on the Prolific platform (www.prolific.com) and paid £6.00 per hour. Since we do not know of any previous studies adapting the two-response paradigm with the CRA, we based our sample size choice on previous studies using the two-response paradigm in the reasoning field (e.g., Bago & De Neys, 2017, 2019b). In Experiment 1, we recruited 100 participants (52 females, *M* age = 37.3, *SD* = 13), among which 32% reported High school, 49% reported a bachelor degree, 16% a master degree, and 3% a PhD as their highest education level. In Experiment 2, we recruited 100 participants (47 females, *M* age = 39, *SD* = 12.7), of which one participant did not report their demographic information. Among the remaining participants, 2% reported less than High school, 27.3% reported High school, 45.5% reported a bachelor degree, 19.2% a master degree, and 6.1% a PhD as their highest education level.

2.3. Compound Remotes Associates Test

In the CRA, participants are presented with three cue words (e.g., river, note, account) and have to find a fourth solution word to combine with each of the cue words to form three new meaningful compound words (e.g., river*bank*, *bank*note, *bank* account). We initially selected candidate items in Bowden and Jung-Beeman (2003), and we created additional items. The newly created items enabled us to cover the whole range of item difficulty, while taking into account the position of the solution word within the created compound words, which could either be at the front, back or a combination of both (i.e., mixed). This preselection resulted in a list of 54 items, of which 17 were new items. We conducted an independent online pilot study using the Prolific platform (n = 50, 25 females, *M* age = 38.8, *SD* = 13.9) to ensure the validity and reliability of our item selection by minimizing the occurrence of alternative correct or incorrect responses. Unlike previous studies using the CRA, when multiple participants provided an alternative yet correct answer, we included it as a valid alternative solution. The pilot study also enabled us to ensure that the items covered a full range of difficulty (Easy: 98–80% accuracy, Medium: 64–46%, Hard: 26–6%). This process resulted in a final selection of 24 items (*M* accuracy = 54.8%, *SD* = 29, range = 6%–98%). Among our items, 9 were front-positioned, 6 were back-positioned, and 9 were mixed-positioned items. The full item list with their solutions can be found in Appendix A. For each trial, participants were presented with the three cue words and had to type their response before pressing 'Enter' to validate it. Appendix B details how we accounted for orthographic and typing mistakes, as well as alternative correct responses.

In addition, although we did not preregister this criterion, we had to exclude one participant in Experiment 1, because they consistently gave inappropriate responses to the CRA problems (e.g., answering "basket" or "basketfootvolley" to the problem "basket/foot/volley").

2.4. Animal Fluency Task

We used the animal fluency task (Ardila et al., 2006), as it is the most widely used to estimate semantic networks (Christensen & Kenett, 2023). Participants had 3 minutes to generate as many animal names as possible. Participants had to type their response before pressing 'Enter' for each new response and were reminded to keep working until the time was over. This task was only used in Experiment 2. Appendix C provides the complete instructions for the animal fluency task.

2.5. Cognitive Load Task

To minimize deliberation in the initial stage of the two-response paradigm, we used the dot memorization task (Miyake et al., 2001). This task has been shown to effectively burden executive resources in verbal reasoning (e.g., De Neys & Verschueren, 2006; Verschueren et al., 2004). In Experiment 1, the participants saw a 3 x 3 grid with four crosses before each CRA problem in the initial stage of the paradigm (see Section 2.8.). After their first response, participants had to select the correct pattern among four different load matrices. They then received feedback about whether they selected the to-be-memorized matrix. The load task was only present during the initial response stage, where deliberation was minimized.

In Experiment 2, we used the same load procedure as in Experiment 1, using a more challenging five dot pattern in a 4 x 4 grid (e.g., Bialek & De Neys, 2017; Trémolière & Bonnefon, 2014). This increased load has been shown to further burden the executive resources of participants compared to the simpler load pattern used in Experiment 1 (Trémolière et al., 2012).

2.6. Deadline Calibration in the Initial Response Stage

2.6.1. Experiment 1

To calibrate an appropriate deadline for the initial stage of the two-response paradigm in Experiment 1, we ran a traditional "one-response" version of the experiment (e.g., Bago et al., 2021). We used the same material as in the two main experiments, but participants only had to give a single answer without deadline or load. We recruited an additional online independent sample of 51 native English speakers on the Prolific platform (26 females, *M* age = 37.2, *SD* = 13.1). Similar to Experiment 1, we excluded 2 participants who consistently misunderstood the CRA task.

Participants took on average 13.5 s (*SD* = 11.2 s) to give a correct answer in the CRA in the one-response pre-test. The first quartile of the reaction times for the correct responses was 7.5 s. Based on this result, we rounded up this value to the nearest integer to give participants some minimal leeway and set the deadline at 8 s in Experiment 1. The screen turned yellow 2 s before the deadline to remind participants of the incoming deadline and urge them to type their response.

To test whether participants were under time pressure in the initial stage of the two-response paradigm, we contrasted reaction times between the initial stage of Experiment 1 (*M* = 5 s, *SD* = 0.8 s) and the one-response pre-test (*M* = 22.6 s, *SD* = 14.7 s). Because responses in Experiment 1 were subject to an 8 s deadline, we used a Bayesian censored (Tobit) log-normal mixed-effects model, treating missed-deadline trials as right-censored rather than excluding them. Results showed that responses in Experiment 1 were about 60% faster than in the one-response pre-test, posterior reaction time ratio = 0.40, 95% CrI [0.36, 0.44], *pd* = 100%.

2.6.2. Experiment 2

In Experiment 2, we used an even more stringent deadline to further minimize the possibility that participants engaged in deliberation during the initial response stage. The new deadline was calibrated based on the average initial response time from Experiment 1. For Experiment 2, we consequently further decreased the deadline to 6 seconds, with the screen turning yellow at 5 seconds to remind participants to type their response.

To test whether the time pressure had increased between Experiment 1 and Experiment 2, we contrasted the reaction times between the initial stage of Experiment 1 (M = 5 s, SD = 0.8 s) and Experiment 2 (M = 4 s, SD = 0.5). Once again, we used a Bayesian censored (Tobit) log-normal mixed-effects model, treating missed-deadline trials in the two experiments as right-censored rather than excluding them. Results showed that reaction times decreased by 17% in the initial response stage of Experiment 2 compared to Experiment 1, posterior reaction time ratio = 0.83, 95% CrI [0.79, 0.87], pd = 100%.

2.7.Final Response Confound

The one-response pre-test also allowed us to investigate whether providing two responses to the same problem in the two-response paradigm biased participants' answers. More specifically, participants might aim to maintain consistency between their initial and final responses. This consistency could potentially hinder them from correcting initial incorrect responses, artificially reducing the accuracy of the final response stage. Conversely, it is also possible that participants' accuracy in the final response stage of the paradigm could be boosted by having already responded once to the same problem in the initial stage.

To rule out these possibilities, we contrasted the accuracy of the final response stage of Experiment 1 (M = 65.2, SD = 15.3) and Experiment 2 (M = 66.7, SD = 18.7) with the accuracy of the one-response pre-test (M = 62.5, SD = 12.8) using mixed-effects logistic regressions including random intercepts for participants and items. There was no significant difference between the one-response pre-test and Experiment 1 (p = .76) nor Experiment 2 (p = .79), suggesting that the repetition of the problems in the two-response paradigm did not influence performance in the final response stage compared to the one-response pre-test.

2.8. Two-Response Paradigm Procedure

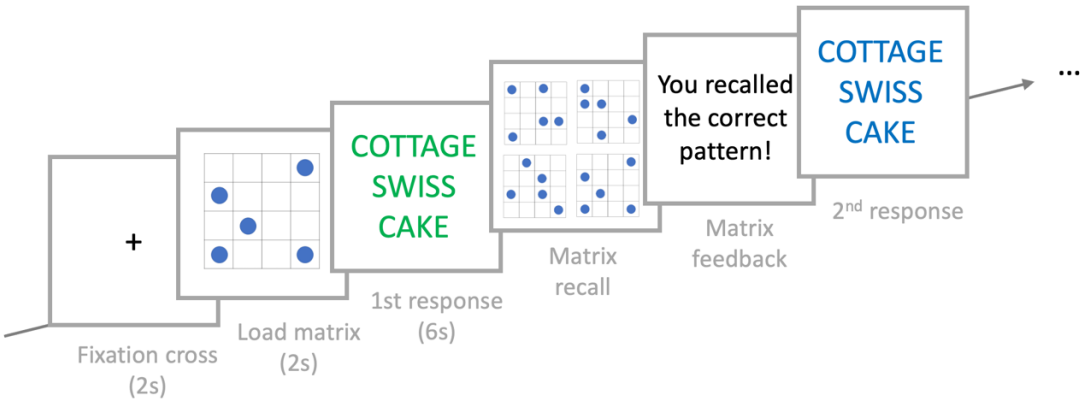

Figure 1. Time course of a two-response CRA trial (Experiment 2).

Appendix C provides the complete instructions and training procedure for the two-response paradigm adaptation of the CRA. The general procedure was similar to Bago and De Neys (2017). The experiments were run online on the Qualtrics platform. A fixation cross was displayed for 2 s at the beginning of each trial. Subsequently, the load matrix was presented for 2 seconds. Participants were then required to solve the CRA problem for the first time, with an allotted

response time of 8 seconds in Experiment 1 (6 seconds in Experiment 2). After 6 seconds (5 seconds, respectively), the screen's background turned yellow as a reminder of the impending deadline. In cases where participants failed to respond before the deadline, they were prompted to provide an answer within the deadline during subsequent initial trials.

Following their initial response, participants were also required to indicate their response confidence by sliding a cursor with their mouse on a horizontal visual analog scale that gradually transitioned from red (representing low confidence = 0) to green (representing high confidence = 100). Subsequently, following Stuyck et al. (2022), participants were tasked with reporting whether they had arrived at a solution with insight or not, choosing between two options: "With Aha!" or "Without Aha!" The instructions explicitly defined each solution type to ensure participants' understanding, following Stuyck et al. (2022; see Appendix C). Following this, participants had to select the correct, to-be-remembered load matrix from among four different matrices. If they were unable to identify the correct load matrix, they were instructed to memorize the pattern correctly in subsequent trials. Note that the confidence and insight questions were added for exploratory purposes (see Appendix D for the results).

The CRA problem was then presented a second time, and participants could give their final response without any deadline or concurrent load. Following this, the participants had to indicate their confidence in their final response and whether they reached it through insight or not. In both response stages, if participants did not provide an answer for the CRA problem, they were not required to indicate response confidence or solution type.

The answer options appeared in green during the initial response stage and switched to blue during the final response stage, serving as a visual cue to remind participants of the question stage they were in. Additionally, a reminder sentence was placed beneath each question: "Please indicate

your very first, intuitive answer" and "Please provide your final answer", respectively. The full procedure is depicted in Fig. 1.

2.9. Trial Exclusion

Following our preregistration criteria, we excluded trials when participants failed to provide an initial response within the deadline or did not recognize the correct load matrix because for these trials, we cannot be sure that participants did not deliberate in the initial response stage. This accounted for 20.1% of the trials in Experiment 1 and 36.7% of the trials in Experiment 2. On average, each participant contributed a total of 19.2 trials in Experiment 1 (SD = 3.3) and 15.2 trials in Experiment 2 (SD = 5.8) out of 24 trials.

2.10. Mixed-effects Models

We used (generalized) linear mixed models to analyze the CRA data, which process trial-by-trial data while accounting for both participant and item variability. Appendix E outlines the method used to find the optimal random structure for each analysis, how we assessed significance across our models, and the contrast coding scheme we used based on the specific needs of each analysis.

2.11. Semantic Similarity Measure

To compute semantic similarity, we used the "*all-mpnet-base-v2*" model from the python package "*sentence-transformers"* (Reimers & Gurevych, 2019). This transformer model captures textual semantic information by ensuring that words or sentences that are semantically close to one another will also be close in the internal vector space of the model. Specifically, each word (or

group of words) is transformed into a vector representation (or word embedding) in a 768-dimensional vector space.

This allows us to compute semantic similarity measures to assess how close words are to one another in terms of meaning. Specifically, we transformed the three cue words (as a unique string) and the solution word(s) into embeddings and then used cosine similarity to compute how close the two resulting vectors were. Cosine similarity is computed as follows:

$$sim(u, v) = \frac{u \cdot v}{\|u\| \times \|v\|} \quad (1)$$

where *u* and *v* are the two vectors we want to compare.

2.12. Associative Thinking and Semantic Memory Structure Measures

2.12.1. Data preprocessing

Appendix B details how we pre-processed the animal fluency data to run our associative thinking analyses in Experiment 2 using the *SemNetCleaner* package in R as described in Christensen and Kenett (2023).

2.12.2. Forward Flow

To compute the forward flow measure for each participant, we used the same transformer model for semantic embedding on the animal names provided by the participants in the additional animal fluency task. We then computed forward flow over the chain of concepts generated by each participant, as shown in Equation 2, following Gray et al. (2019):

$$\frac{1}{n-1}\sum_{i=2}^{n}\left(\frac{\sum_{j=1}^{i-1} D_{i,j}}{i-1}\right) \quad (2)$$

where $D_{i,j}$ is the semantic distance between two words embeddings computed as 1 – cosine similarity (from Equation 1) and *n* is the total number of words generated by the participant in the animal fluency task. Equation 2 thus computes, for each newly generated animal, the average semantic distance between this animal and every previously generated animal. It then computes the mean of these averages to yield a synthetic measure of how efficiently the participant explores the semantic space.

2.12.3. Semantic Network Estimations

To compute the semantic networks at the group level, we used the pipeline described in the SemNa package tutorial (Christensen & Kenett, 2023). Because semantic network estimation from fluency data involves substantial researcher degrees of freedom and there is no consensus on a single "best" method (Zemla & Austerweil, 2018), we adopted a robustness strategy and applied the four estimation methods available in the package: the Correlation-Based Network method, the Naive Random Walk method, the Pathfinder Network method, and the Community Network method (see Appendix F for construction details).

In line with Luchini et al. (2023), we then estimated semantic networks as a function of intuitive performance in the CRA by using a median split to contrast the High "11" group vs the Low "11" group semantic networks. Next, we tested whether the generated networks were significantly different from random networks with the same number of nodes and edges for each network metric. This led us to exclude the Community Network method, for which the metrics consistently did not differ from those of randomly generated networks.

We then compared the networks using case-wise bootstrap analysis. Because semantic network metrics can depend on the estimation procedure and results did not always converge across methods, we aggregated method-specific standardized effect sizes to summarize the overall direction and magnitude of group differences across approaches, using meta-analytic procedures for dependent effect sizes (Borenstein et al., 2021). Appendix F details the complete procedure for the semantic network analysis.

3. Results

3.1. Quantifying Correct Intuitive Versus Deliberate Performance

Fig. 2a provides a summary of the initial and final accuracies for the two experiments. The average accuracy was significantly lower in the initial response stage (Experiment 1: $M = 48.3\%$, $SD = 14.5$; Experiment 2: $M = 49.8\%$, $SD = 22.8$) than in the final response stage (Experiment 1: $M = 65.2\%$, $SD = 15.3$; Experiment 2: $M = 66.7\%$, $SD = 18.7$). A mixed-effects logistic regression, including random intercepts for items and random intercepts and slopes for response stage for participants, revealed a significant effect of response stage on accuracy, $OR = 1.89$, $p < .001$, Cohen's $d = 0.35$, 95% CI [0.31, 0.39], but no significant effect of experiment ($p = .74$) nor of the response stage by experiment interaction ($p = .17$). Thus, the results consistently show that although deliberation helped, participants still managed to give a high number of correct responses in the initial response stage where deliberation was minimized.

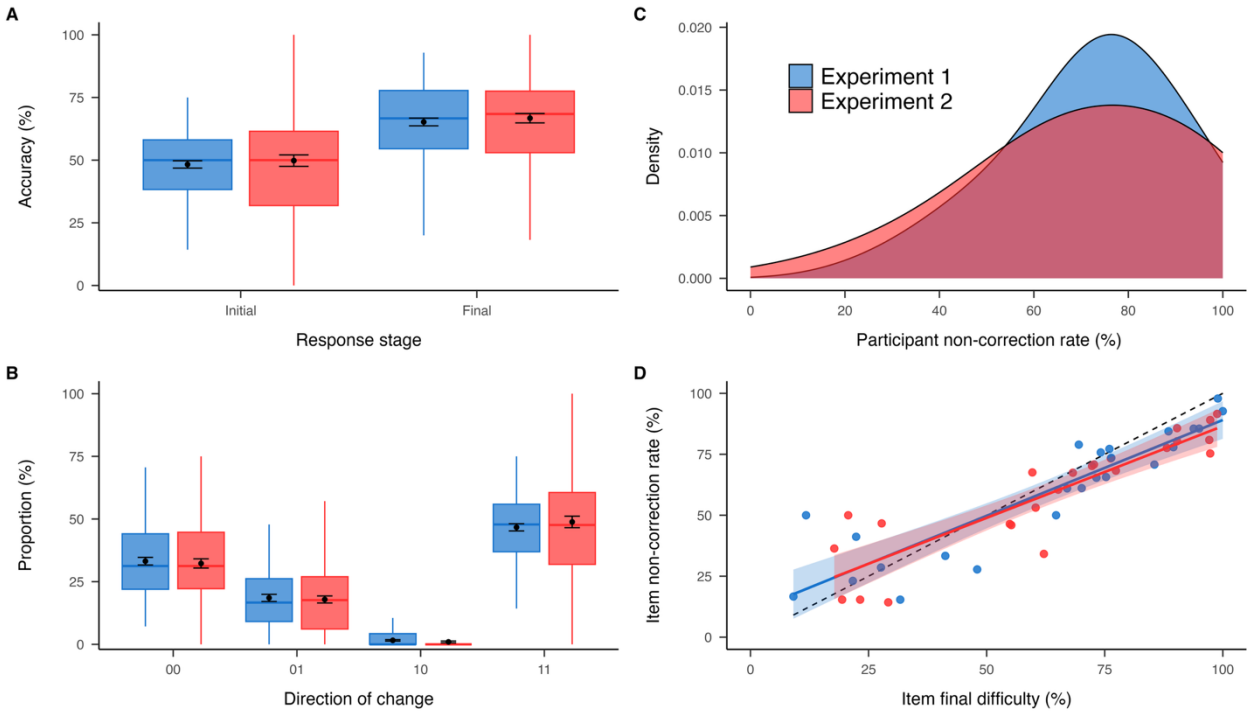

Figure 2. By-subject and by-item performance analyses across response stages in Experiment 1 and Experiment 2. **a)** Response accuracy as a function of response stage, by subject. **b)** Proportion of each direction-of-change category, by subject: "00" = incorrect initial and incorrect final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct initial and correct final response. In the boxplots, the lower and upper hinges correspond to the first and third quartiles, the middle line shows the median, and the whiskers extend to the smallest and largest values no further than 1.5 times the interquartile range. Overlaid black dots represent the mean, and black error bars indicate standard errors of the mean. **c)** Distribution of the average non-correction rate across participants, by subject. **d)** Relationship between final response accuracy and non-correction rate at the item level. Solid lines with shaded ribbons show linear regression fits with 95% confidence intervals; the dashed line indicates the identity line.

To better understand how deliberation affected participants' responses, we performed a direction of change analysis (Bago & De Neys, 2017). In each trial, participants can give either a correct response (accuracy = 1) or an incorrect response (accuracy = 0) in the two response stages (initial and final). This yields four possible directions 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). Fig. 2b summarizes those results.

To obtain a straightforward measure of how “11” responses compared to “01” responses, we computed the mean “non-correction rate” across our participants. This rate, expressed as the proportion 11/(11+01) (Bago & De Neys, 2017), indicates the proportion of correct responses in the final response stage that were already correct in the initial response stage. A high non-correction rate indicates that most correct responses were already accurate in the initial response stage, implying that participants did not require deliberation to arrive at the correct answer. The mean non-correction rate was 72.7% (*SD* = 23.2) in Experiment 1 and 71.4% (*SD* = 23.2) in Experiment 2 (Fig. 2c).

To statistically assess whether there was a higher occurrence of sound intuitive thinking (“11”) as opposed to correct deliberate responses following an initial incorrect response (“01”), we focused exclusively on final correct responses, excluding “00” and “10” trials. We then built a mixed-effects logistic regression model, which included random intercepts for both participants and items, with the experiment as a fixed effect. We used a dummy variable to code whether a trial was a “01” response (coded as 0) or a “11” response (coded as 1). The effect of experiment was not significant ($p = .36$). Importantly, the estimated non-correction rates from the model were large both for Experiment 1, $M = 66\%$, 95% CI [53, 78] and Experiment 2, $M = 63\%$, 95% CI [50, 76]. Thus, when participants gave a correct answer in the final response stage of the paradigm, they had already given a correct answer in the initial response stage most of the time.

To assess the limits of sound intuitive thinking in the CRA, we examined how item difficulty related to the likelihood of sound intuitive thinking (Fig. 2d). Item difficulty was operationalized as the average accuracy for each item in the final response stage, providing an estimate of difficulty

once participants were allowed to deliberate. We fitted linear regression models predicting the by-item non-correction rate from by-item difficulty in each experiment. Results showed that item difficulty was a strong positive predictor of the non-correction rate in both Experiment 1, $b = 0.79$, $t(22) = 9.84$, $p < .001$, and Experiment 2, $b = 0.75$, $t(22) = 9.10$, $p < .001$ (see Appendix G for the full models). This suggests that as items became more challenging, the occurrence of sound intuitive thinking ("11" responses) decreased relative to deliberate correction following an initial error ("01" responses), which highlights a boundary condition for intuitive processing. Consistently, item difficulty was also very strongly correlated with slower reaction times in the final response stage, Experiment 1: $r(22) = -.89$, $p < .001$, 95% CI [−.95, −.75]; Experiment 2: $r(22) = -.93$, $p < .001$, 95% CI [−.97, −.85], indicating that harder items tended to elicit longer deliberation before the final answer.

However, sound intuitive thinking remained possible even at the highest levels of difficulty: for the hardest decile of items in each experiment (final-stage accuracy = 14.2% in Experiment 1; 19.3% in Experiment 2), the models still predicted non-correction rates of 21.6% (empirical = 29.9%) and 25.8% (empirical = 33.9%), respectively. This indicates that although sound intuitive thinking becomes less probable as difficulty increases, it does not disappear entirely.

3.2. Exploratory Confidence Analysis

One may wonder whether participants exhibit intuitive metacognitive sensitivity by reporting lower confidence when giving an incorrect response in the initial stage of the paradigm. The results showed that participants reported higher confidence for correct responses than for incorrect responses. Participants were thus able to recognize when they had not converged on the correct solution in the CRA, even when deliberation was minimized. This strong error sensitivity can be

explained by the fact that an incorrect candidate response fails to form three valid compound words from the cues, leading to reduced experienced confidence.

A related question is whether lower confidence in the initial stage predicts higher accuracy in the final stage. Such a pattern would suggest the engagement of deliberative thinking when initial certainty is low, as previously observed in the reasoning literature (e.g., Bago & De Neys, 2017). To test this, we examined whether initial confidence on incorrect responses predicted final accuracy by comparing "00" and "01" responses (thus controlling for initial accuracy). Our results clearly show that, in both experiments, initial confidence did not predict final accuracy. Full details of these supplementary confidence analyses are provided in Appendix H for the interested reader.

3.3.Associative Processes Account for Intuitive Performance

Our results point to a clear role for sound intuitive thinking in creative idea generation. To clarify how such intuitions arise, we examined associative processes underlying convergent thinking at both the item and participant levels.

3.3.1. Items' Semantic Search Space Correlates with Sound Intuitive Thinking

At the item level, we tested whether semantic similarity between the three cue words and their solution could explain why some problems more readily elicit correct intuitions than others in CRA items (e.g.,"extinguisher, truck, camp" is closer in meaning to its solution "fire," while "baby, spring, cap" is more distant from its solution "shower" and thus harder).

To compute the semantic similarity between the string of three cue words and the solution word for each item, we employed a state-of-the-art transformer model (see Section 2.11.). Results indicated that there was a large and positive correlation between semantic similarity and the "11"

response rate in Experiment 1, $r(22) = .54$, $p = .006$, 95% CI [.18, .78], as well as in Experiment 2, $r(22) = .54$, $p = .007$, 95% CI [.17, .77]. However, this was not the case for "01" response in either Experiment 1, $r(22) = .05$, $p = .83$, 95% CI [-.36, .44] or in Experiment 2, $r(22) = -.01$, $p = .96$, 95% CI [-.41, .39]. In line with these findings, there was also a positive correlation between semantic similarity and the by-item non-correction rate in Experiment 1, $r(22) = .43$, $p = .035$, 95% CI [.04, .71], and in Experiment 2, $r(22) = .45$, $p = .026$, 95% CI [.06, .73]. Therefore, as shown in Fig. 3a, items with higher semantic similarity between the cue words and the solution word were associated with a higher likelihood of sound intuitive thinking ("11" responses) compared to deliberate correction following an initial failure ("01" responses). Put differently, semantic similarity predicted the likelihood of sound intuitive thinking but not sound deliberate thinking in our experiments. This pattern of results is consistent with the idea that associative processes are more likely to be at play during intuitive rather than deliberate reasoning.

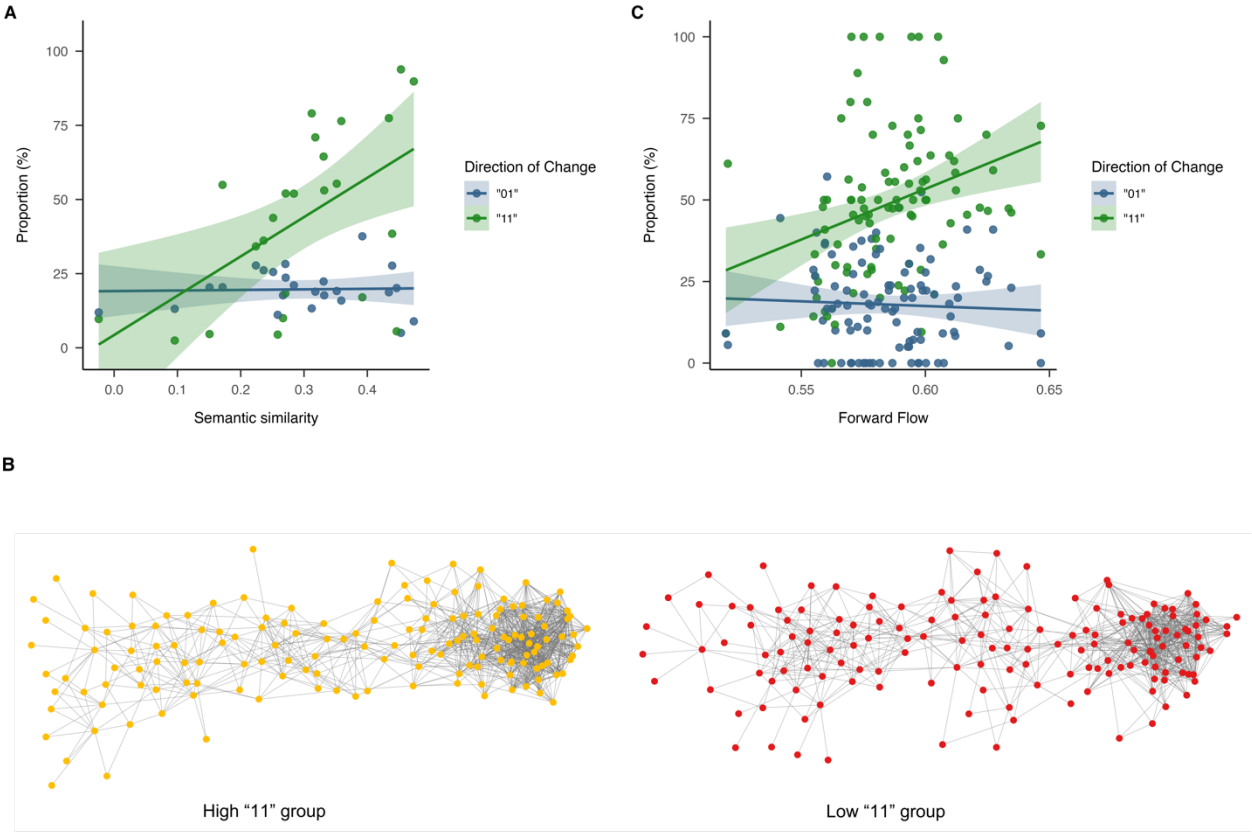

Figure 3. Associative processes at the item and participant levels relate to intuitive performance in the CRA. **a)** Relationship between the semantic similarity of the cue words' string to the solution word and the proportion of "11" direction of change category and "01" direction of change category at the item level, averaged over the two experiments. Lines show linear regression fits with shaded areas representing the 95% confidence interval. **b)** Example visualization of the semantic networks of the High "11" group (orange, left panel) and Low "11" group (red, right panel) in Experiment 2 using the Pathfinder network method. Circles (nodes) represent concepts, and lines (edges) represent the strength of the semantic associations between the concepts for the two groups, where shorter lines reflect stronger associations. **c)** Relationship between forward flow in the animal fluency task and the proportion of "11" direction of change category and "01" direction of change category at the participant level in Experiment 2. Lines show linear regression fits with shaded areas representing the 95% confidence interval. These results demonstrate that higher intuitive performance in the CRA is associated with items characterized by a smaller semantic search space and participants showing a more flexible and efficient semantic memory structure.

3.3.2. Interindividual Differences in Semantic Memory Structure Relate to Higher Intuitive Performance

At the participant level, we tested whether differences in the semantic memory structure could affect their (intuitive) abilities in the CRA task. Prior research indicates that creative individuals tend to have more flexible and interconnected semantic networks, which reduce the need for controlled, deliberate search (Kenett et al., 2014, 2018; Luchini et al., 2023). To test this, we used a verbal fluency task in Experiment 2, in which participants have to generate as many animal names as possible in three minutes. This task allowed us to map participants' semantic memory structure by estimating how closely concepts were interconnected. We then compared networks for individuals with higher versus lower intuitive performance on the CRA.

To contrast participants with higher and lower intuitive performance, we conducted a median split on the proportion of "11" responses in the CRA in Experiment 2. We then estimated semantic memory networks separately for the two resulting groups using three different estimation techniques: the naïve random walk, the pathfinder network, and the correlation-based methods (see Section 2.12.3.). Fig. 3b shows an example of estimated networks for the two groups. We compared the estimated networks on three critical network metrics that have been shown to correlate with creativity (e.g., Benedek et al., 2017; Kenett et al., 2014). The first, Average Shortest Path Length (ASPL), quantifies the average minimum steps necessary to connect any two nodes in the network. A higher ASPL indicates a more spread-out network structure. The second, Clustering Coefficient (CC), measures the likelihood that two adjacent nodes of the same node are also directly connected, reflecting local interconnectivity. The third metric, modularity (Q), assesses the extent to which the network is segmented into smaller, distinct sub-networks, which can correspond to semantic

categories (e.g., farm animals vs. pets). A higher Q indicates greater separation between these sub-networks.

To compare the network structures of the two groups statistically, we conducted independent *t*-tests using case-wise bootstrap analysis. We report the results for each network estimation method separately in Appendix F. To summarize these results, we computed pooled standardized effect sizes while accounting for the dependence between these measures (see Section 2.12.3.). Participants with higher intuitive performance exhibited a semantic memory structure characterized by less dispersion (i.e., lower ASPL; Hedges' g = -0.25, 95% CI [-0.33, -0.17], Z = -6.01, $p < .001$), greater interconnectedness (i.e., higher CC; Hedges' g = 0.61, 95% CI [0.52, 0.7], Z = 13.62, $p < .001$), and a reduced number of sub-networks (i.e., lower Q; Hedges' g = -0.66, 95% CI [-0.75, -0.57], Z = -14.61, $p < .001$).

Overall, compared to the Low “11” group, the High “11” group's semantic memory network had shorter mean path lengths (lower ASPL), showed significantly greater connectivity (higher CC), and contained fewer distinct sub-networks (lower Q). Taken together, these results suggest that sound intuitive reasoners had a denser and more interconnected semantic memory structure, likely allowing them to reach a solution to the CRA problem more easily in the initial, intuitive response stage (Fig. 3b).

3.3.3. Individual Exploration of Semantic Space Relates to Intuitive Performance

To further validate our group-level analysis, we investigated whether the effectiveness with which each participant navigated the semantic space in the animal fluency task was related to their intuitive performance in the CRA in Experiment 2. We computed the forward flow for each participant using the verbal fluency list generated by each participant. This metric captures the

average semantic progression of a chain of thoughts over time by assessing how closely each new concept aligns semantically with the bulk of previously generated concepts, yielding a synthetic measure of how well one explores the semantic space (Gray et al., 2019). First, using the same transformer model as in the semantic search space analysis, we quantified the average semantic distance between each newly generated word in the verbal fluency task and all the preceding words generated by the participant. We then calculated the mean of these individual averages to obtain the forward flow for each participant (see Section 2.12.2.).

Fig. 3c summarizes the results. We observed a moderate correlation between forward flow and the proportion of "11" responses, r(97) = .32, p = .001, 95% CI [.13, .48], but not the proportion of "01" responses, r(97) = -.05, p = .64, 95% CI [-.24, .15]. To address potential confounding factors, we conducted a multiple linear regression using the "11" proportion as the dependent variable and included forward flow, verbal fluency (i.e., the number of words produced by participants), and the participant's education level as independent variables. Even after adjusting for these additional variables, the effect of forward flow remained the only statistically significant predictor of "11" response proportion, $b = 6.01$, $t(94) = 2.35$, $p = .02$ (see Appendix G for the full model). These results are in line with our group-level analysis of semantic memory structure, suggesting that participants who explored the semantic space more efficiently (during the verbal fluency task) also tended to show a better intuitive solution performance in the CRA.

4. Discussion

In a series of two experiments, we used a two-response paradigm where participants were required to provide their intuitive, initial responses under both concurrent cognitive load and time pressure, to ascertain whether participants were able to give the correct response intuitively in the

CRA. Remarkably, the results revealed that a majority of accurate responses were already correct right from the initial intuitive stage where deliberation was minimized. This suggests that intuitive processes may play a larger role in producing the correct response in the CRA than previously assumed.

What underlies this high prevalence of correct intuitive responses? Our findings point to two key factors. First, we observed a strong positive correlation between the semantic similarity of cue words to the solution word and the proportion of correct intuitive responses: problems with a smaller semantic search space were more likely to be solved intuitively. Second, interindividual differences in associative thinking were also significantly related to intuitive performance. Through an additional semantic fluency task, we estimated the structure of participants' semantic memory networks in Experiment 2. We found that higher intuitive performance in the CRA was linked to a more efficient and flexible semantic memory structure at the group level. Specifically, this structure exhibited characteristics of a "small-world" network—high local connectivity, shorter average path lengths, and lower modularity—allowing concepts that are more distantly related to be connected with greater efficiency and flexibility (He et al., 2020; Kenett et al., 2018). This result replicates prior work showing that convergent thinking performance measured in standard one-response CRA tasks is related to a more flexible semantic memory structure (Luchini et al., 2023) and extends it by showing that similar structural properties are associated with intuitive performance under strong time and resource constraints. To further validate these findings, we modeled the efficiency of semantic exploration by computing a forward flow metric at the individual level (Gray et al., 2019). We found that the efficiency of participants' exploration of the semantic space was positively linked to their intuitive performance in the CRA. Taken together, these results provide compelling

evidence that correct responses in convergent thinking tasks may often rely on associative processes that unfold effortlessly within semantic memory.

Overall, these findings suggest that participants may sometimes converge on the correct response intuitively by relying on associative mechanisms, without the need for further controlled, deliberate processes in the CRA. They also align with results on insight, which show that problems solved with insight can often be solved intuitively (Stuyck et al., 2022). This pattern, however, was not observed in our experiments: although insight was consistently associated with higher accuracy, we found no evidence that cognitive load or response deadlines differentially affected insight versus non-insight solutions (see Appendix D for the full results of this exploratory analysis). However, we do not argue that all creative tasks can be performed purely intuitively. Even within the CRA, our data show that when the semantic distance between the solution word and the cue words increases, deliberation becomes more critical. In such cases, deliberate thought enables a broader exploration of candidate solutions.

Thus, our findings suggest a nuanced understanding of creative thinking, where the necessity of deliberation depends on both task-related factors, such as the semantic search space, and interindividual differences. For instance, individuals with a more efficient semantic memory structure may rely less on costly deliberative thinking, aligning with neuroimaging studies that suggest greater neural efficiency during creative tasks for more creative people (Chen et al., 2025; Chrysikou et al., 2020; Herault et al., 2024; Japardi et al., 2018). Similarly, we can predict that tasks requiring the establishment of more distant connections, or which strongly elicit an incorrect, intuitive answer, will benefit more from deliberation (e.g., the egg task; Camarda et al., 2024), highlighting the need to generalize these findings beyond CRA to other forms of creative generation. In this respect, our findings align with recent theories that offer a more nuanced view

of the interplay between intuitive and deliberate processes in creativity. The "distance-dependent representation activation mode" hypothesis (D-DRAM; Volle, 2018), for instance, suggests a dynamic balance between associative and controlled processes, influenced by task demands, the phase of the creative process, and interindividual differences. However, while the D-DRAM hypothesis emphasizes the role of controlled processes in the evaluative phase of creativity—such as inhibiting common ideas—our results point toward a complementary perspective: the evaluative phase of creativity can still occur effectively even when deliberation is minimized.

Interestingly, the findings from our experiments nicely align with those observed in the logical reasoning field, where people are often able to produce correct responses intuitively (e.g., Bago & De Neys, 2017, 2019b; Thompson & Johnson, 2014). This automatic access to logical and probabilistic rules has led to the hypothesis of "logical intuitions" (De Neys, 2012; Handley & Trippas, 2015; Thompson & Newman, 2017). However, the precise nature of these intuitions remains a subject of ongoing debate in the reasoning field (De Neys, 2023). Recent experimental work has shown that these intuitions may, in fact, not rely on logical rules and operations, but rather on external surface cues that only align with logic (e.g., Ghasemi et al., 2022, 2023; Meyer-Grant et al., 2023). Importantly, unlike in the reasoning domain, where the exact nature of intuitions is more difficult to characterize, our computational and semantic analyses allow us to constrain plausible mechanisms underlying sound intuitive thinking in creative problem-solving. Our findings suggest that, in this context, intuition operates through the rapid associative activation between concepts, providing a mechanistic explanation for sound intuitive thinking. At the same time, it also allows us to specify their boundary condition, as the likelihood of sound intuitive thinking decreases as the semantic distance between the solution word and the problem cues increases.

Although we did not include any direct measure of cognitive capacity (e.g., Raven's Matrices) in this study, future research could examine how well such a measure predicts intuitive performance in the CRA beyond our associative thinking and semantic memory structure measures. This question is particularly relevant because convergent thinking tasks such as the CRA may assess intelligence rather than creativity per se (Chuderski & Jastrzębski, 2018). Such a finding would also align with emerging evidence from the reasoning literature suggesting that cognitive capacity would be a better predictor of sound intuitive thinking than correct deliberation in reasoning (Raoelison et al., 2020; Thompson et al., 2018).

To avoid any misinterpretation, we emphasize that our use of the dual-process framework and the labels "intuitive" and "deliberative" is primarily intended as a pragmatic tool for communicating among scholars. In this study, "intuition" was defined operationally: we combined task instructions, time pressure, and concurrent load to reduce the engagement of cognitive resources during the initial response stage of our paradigm. However, the two-response paradigm cannot, by itself, resolve the broader debate about whether a given response is truly "intuitive" or "deliberative" in nature. Accordingly, our results do not imply that the distinction between intuitive and deliberate thinking is qualitative rather than quantitative. The current data are equally compatible with a single-process account in which intuition and deliberation are conceived as opposite ends of a processing continuum (De Neys, 2021; Hayes et al., 2018). We therefore do not take a position in this broader debate. Instead, our key point is that the intuitive end of the processing continuum may be more capable than traditionally assumed, and that converging on the correct response in the CRA does not necessarily require engaging additional cognitive resources or extended deliberation.

Although we designed our study to cover a broad range of difficulty using items adapted from Bowden and Jung-Beeman (2003), a potential concern is that strong intuitive performance may be due to selecting relatively easier items than those typically found in the literature. To address this, we compared our results with other CRA studies and found that the combined average accuracy in the final response stage (61%) falls within the typical range reported in the literature (42%–80%; see Appendix I). Thus, our items are neither exceptionally easy nor difficult.

Another potential criticism may be that the presence of sound intuitive thinking in our studies was due to the fact that our methodology did not sufficiently hamper deliberation in the initial response stage. It is worth noting that we combined three established procedures (instructions, time constraints, and cognitive load) to ensure that reasoners could not engage in deliberation. These methods have all been demonstrated to be effective in limiting deliberation. Furthermore, the chosen time limit was calibrated using a pilot study in Experiment 1 (see Section 2.6.1.). In Experiment 2, we introduced an even more stringent cognitive load task and time limit to further minimize the possibility of engaging in deliberation. The high number of excluded trials (20.1% in Experiment 1; 36.7% in Experiment 2) indicates that using a more demanding deadline or load would have created practical and statistical issues, such as selection effects at the subject level (Bouwmeester et al., 2017). However, we acknowledge that it remains possible some participants engaged in deliberation during the initial response stage, despite our methodological constraints.

One possible concern is that participants may have strategically safeguarded their performance on the CRA, by prioritizing the CRA items over responding before the deadline or succeeding in the cognitive load task. If so, this strategy would likely be more common on difficult CRA items, where cognitive demand is greater. Such behavior could artificially inflate apparent intuitive performance in the initial stage due to trial exclusions, as previously noted (e.g., Stuyck

et al., 2022). More precisely, if excluded trials disproportionately involved difficult items, their removal could lead to an overestimation of correct intuitive responses through selective truncation of errors on harder items. To address this concern, we conducted additional analyses in which missed-deadline trials were conservatively recoded as incorrect. Although the prevalence of sound intuitive thinking decreased slightly under this stricter approach, it remained very high (see Appendix J). Thus, our main findings are robust even under conservative assumptions about trial exclusions.

Note that one could still argue that the response deadline and the cognitive load we used were not challenging enough and that the correct initial responses we observed would disappear "with just a little more load or time pressure". However, such arguments make dual-process theories difficult to falsify at this point, as any evidence for sound intuitive thinking can always be dismissed by arguing that the methods left space for deliberation. From a more theoretical perspective, the issue lies in the fact that dual-process theories are not fully specified (Kruglanski, 2013). The framework typically suggests that System 2 operates more slowly and requires more cognitive resources than System 1, yet it does not provide a clear a priori criterion for determining whether a process is intuitive or deliberate (e.g., requires at least x amount of time or cognitive load; De Neys, 2023).

Conversely, we cannot entirely rule out the possibility that participants relied on intuitive rather than deliberative processes during the final response stage. Although our data indicated that response times in the final stage of the two-response paradigm were not shorter than in the one-response pre-test, future research could investigate whether performance in the final stage might be further enhanced by explicitly promoting deliberation (e.g., through incentives or instructional manipulations).

To conclude, the present paper shows that convergent thinking does not necessarily require extended deliberation: under strong time and resource constraints, correct solutions can emerge through associative processes supported by semantic memory.

5. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI's GPT-4 model (2023) to proofread the manuscript for grammatical and stylistic improvements. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Acknowledgements: This research was supported by a grant from the Agence Nationale de la Recherche, France (ANR-23-CE28-0004-01). We would like to thank Dr. Esther Boissin and Dr. Aikaterini Voudouri for their contributions to the initial discussion and design of our experiments.

Ethics statement: All experimental procedures were conducted in accordance with relevant laws and institutional guidelines and approved by the institutional ethics committee (CER U-Paris; approval date: December 17, 2019). The privacy rights of all human subjects were strictly observed, and informed consent was obtained from all participants involved in the study.

Credit author statement: Jérémie Beucler: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Wim De Neys**: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review & editing.

References

Allen, A. P., & Thomas, K. E. (2011). A dual process account of creative thinking. *Creativity Research Journal*, *23*(2), 109–118. https://doi.org/10.1080/10400419.2011.571183

Ardila, A., Ostrosky-Solís, F., & Bernal, B. (2006). Cognitive testing toward the future: The example of Semantic Verbal Fluency (ANIMALS). *International Journal of Psychology*, *41*(5), 324–332. https://doi.org/10.1080/00207590500345542

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

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2014). Reasoned connections: A dual-process perspective on creative thought. *Thinking & Reasoning*, *21*(1), 61–75. https://doi.org/10.1080/13546783.2014.895915

Beaty, R. E., & Kenett, Y. N. (2023). Associative thinking at the core of creativity. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2023.04.004

Beaty, R. E., Zeitlen, D. C., Baker, B. S., & Kenett, Y. N. (2021). Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity*, *41*, 100859. https://doi.org/10.1016/j.tsc.2021.100859

Benedek, M., Beaty, R. E., Schacter, D. L., & Kenett, Y. N. (2023). The role of memory in creative ideation. *Nature Reviews Psychology*, *2*(4), 246–257. https://doi.org/10.1038/s44159-023-00158-z

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., & Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: a network science approach. *Thinking & Reasoning*, *23*(2), 158–183. https://doi.org/10.1080/13546783.2016.1278034

Beucler, J., Voudouri, A., & De Neys, W. (2025). Moses illusions, fast and slow. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advance online publication. https://doi.org/10.1037/xlm0001530

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John wiley & sons. https://doi.org/10.1002/9780470743386

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

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 634–639. https://doi.org/10.3758/BF03195543

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

Camarda, A., De Neys, W., Ozkalp-Poincloux, B., Hooge, S., Le Masson, P., Weil, B., & Cassotti, M. (2024). Detecting fixation bias in creative idea generation: Evidence from design novices and experts. Creativity Research Journal, 1-21. https://doi.org/10.1080/10400419.2024.2424620

Cassotti, M., Agogué, M., Camarda, A., Houdé, O., & Borst, G. (2016). Inhibitory Control as a Core Process of Creative Problem Solving and Idea Generation from Childhood to Adulthood. *New directions for child and adolescent development*, *2016*(151), 61–72. https://doi.org/10.1002/cad.20153

Chen, Q., Kenett, Y. N., Cui, Z., Takeuchi, H., Fink, A., Benedek, M., Zeitlen, D. C., Zhuang, K., Lloyd-Cox, J., Kawashima, R., Qiu, J., & Beaty, R. E. (2025). Dynamic switching between

brain networks predicts creative ability. *Communications Biology*, *8*(1), 54. https://doi.org/10.1038/s42003-025-07470-9

Christensen, A. P., & Kenett, Y. N. (2023). Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychological methods*, *28*(4), 860–879. https://doi.org/10.1037/met0000463

Chrysikou, E. G., Jacial, C., Yaden, D. B., van Dam, W., Kaufman, S. B., Conklin, C. J., Wintering, N. A., Abraham, R. E., Jung, R. E., & Newberg, A. B. (2020). Differences in brain activity patterns during creative idea generation between eminent and non-eminent thinkers. *NeuroImage*, *220*, 117011. https://doi.org/10.1016/j.neuroimage.2020.117011

Chuderski, A., & Jastrzębski, J. (2018). Much ado about aha!: Insight problem solving is strongly related to working memory capacity and reasoning ability. *Journal of Experimental Psychology: General, 147*(2), 257–281. https://doi.org/10.1037/xge0000378

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. (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., & 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

Evans, J. S., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Ghasemi, O., Handley, S., Howarth, S., Newman, I. R., & Thompson, V. A. (2022). Logical intuition is not really about logic. *Journal of Experimental Psychology: General, 151*(9), 2009–2028. https://doi.org/10.1037/xge0001179

Ghasemi, O., Handley, S. J., & Howarth, S. (2023). Illusory intuitive inferences: Matching heuristics explain logical intuitions. *Cognition*, *235*, 105417. https://doi.org/10.1016/j.cognition.2023.105417

Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., Huang, L., Kenett, Y. N., & Lewis, K. (2019). "Forward flow": A new measure to quantify free thought and predict creativity. *The American psychologist*, *74*(5), 539–554. https://doi.org/10.1037/amp0000391

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in cognitive sciences*, *6*(12), 517-523. https://doi.org/10.1016/S1364-6613(02)02011-9

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In *Psychology of learning and motivation* (Vol. 62, pp. 33-58). Academic Press. https://doi.org/10.1016/bs.plm.2014.09.002

Hayes, B. K., Stephens, R. G., Ngo, J., & Dunn, J. C. (2018). The dimensionality of reasoning: Inductive and deductive inference can be explained by a single process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1333. https://doi.org/10.1037/xlm0000527

He, L., Kenett, Y. N., Zhuang, K., Liu, C., Zeng, R., Yan, T., Huo, T., & Qiu, J. (2020). The relation between semantic memory structure, associative abilities, and verbal and figural creativity. *Thinking & Reasoning*, *27*(2), 268–293. https://doi.org/10.1080/13546783.2020.1819415

Herault, C., Ovando-Tellez, M., Lebuda, I., Kenett, Y. N., Beranger, B., Benedek, M., & Volle, E. (2024). Creative connections: the neural correlates of semantic relatedness are associated with creativity. *Communications biology*, *7*(1), 810. https://doi.org/10.1038/s42003-024-06493-y

Japardi, K., Bookheimer, S., Knudsen, K., Ghahremani, D. G., & Bilder, R. M. (2018). Functional magnetic resonance imaging of divergent and convergent thinking in Big-C creativity. *Neuropsychologia*, *118*(Pt A), 59–67. https://doi.org/10.1016/j.neuropsychologia.2018.02.017

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

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience*, *8*, 407. https://doi.org/10.3389/fnhum.2014.00407

Kenett, Y. N., Levy, O., Kenett, D. Y., Stanley, H. E., Faust, M., & Havlin, S. (2018). Flexibility of thought in high creative individuals represented by percolation analysis. *Proceedings of the National Academy of Sciences*, *115*(5), 867–872. https://doi.org/10.1073/pnas.1717362115

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

Luchini, S., Kenett, Y. N., Zeitlen, D. C., Christensen, A. P., Ellis, D. M., Brewer, G. A., & Beaty, R. E. (2023). Convergent thinking and insight problem solving relate to semantic memory network structure. *Thinking Skills and Creativity*, *48*, 101277. https://doi.org/10.1016/j.tsc.2023.101277

Marko, M., Michalko, D., & Riečanský, I. (2019). Remote associates test: An empirical proof of concept. *Behavior research methods*, *51*(6), 2700-2711. https://doi.org/10.3758/s13428-018-1131-7

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

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220. https://doi.org/10.1037/h0048850

Meyer-Grant, C. G., Cruz, N., Singmann, H., Winiger, S., Goswami, S., Hayes, B. K., & Klauer, K. C. (2023). Are logical intuitions only make-believe? Reexamining the logic-liking effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(8), 1280–1305. https://doi.org/10.1037/xlm0001152

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

OpenAI. (2023). Gpt-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Ovando-Tellez, M., Kenett, Y., Benedek, M., Hills, T., Beranger, B., Lopez-Perseem, A., & Volle, E. (2024). Switching, fast and slow: Deciphering the dynamics of memory search, its brain

connectivity patterns, and its role in creativity. *Research Square*. https://doi.org/10.21203/rs.3.rs-3826172/v1

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

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

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv*. https://doi.org/10.18653/v1/D19-1410

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making, 7*(3), 332-359. https://doi.org/10.1017/S1930297500002291

Rossmann, E., & Fink, A. (2010). Do creative people use shorter associative pathways? *Personality and Individual Differences*, *49*(8), 891–895. https://doi.org/10.1016/j.paid.2010.07.025

Sowden, P. T., Pringle, A., & Gabora, L. (2019). The shifting sands of creative thinking: Connections to dual-process theory. In *Insight and Creativity in Problem Solving* (pp. 40–60). Routledge. https://doi.org/10.4324/9781315144061-3

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

Stuyck, H., Cleeremans, A., & Van den Bussche, E. (2022). Aha! Under pressure: The Aha! Experience is not constrained by cognitive load. *Cognition*, *219*, 104946. https://doi.org/10.1016/j.cognition.2021.104946

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., & Newman, I. R. (2017). Logical intuitions and other conundra for dual process theories. In *Dual process theory 2.0* (pp. 121-136). Routledge. https://doi.org/10.4324/9781315204550

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

Topolinski, S., & Strack, F. (2009). Scanning the "fringe" of consciousness: What is felt and what is not felt in intuitions about semantic coherence. Consciousness and cognition, 18(3), 608-618. https://doi.org/10.1016/j.concog.2008.06.002

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

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. Retrieved from https://escholarship.org/uc/item/7kk1x3qx

Volle, E. (2018). Associative and controlled cognition in divergent thinking: Theoretical, experimental, neuroimaging evidence, and new directions. In R. E. Jung & O. Vartanian (Eds.), *The Cambridge handbook of the neuroscience of creativity* (pp. 333–360). Cambridge University Press. https://doi.org/10.1017/9781316556238.020

Voudouri, A., Bago, B., Borst, G., & De Neys, W. (2023). Reasoning and cognitive control, fast and slow. *Judgment and Decision Making*, *18*, e33. https://doi.org/10.1017/jdm.2023.32

Zemla, J. C., & Austerweil, J. L. (2018). Estimating semantic networks of groups and individuals from fluency data. *Computational brain & behavior*, *1*(1), 36-58. https://doi.org/10.1007/s42113-018-0003-7

Appendices - Intuitive Insight: Fast Associative Processes Drive Sound Creative Thinking

A. CRA Problems

Table A.1. List of the CRA items used in our experiments.

Item number	CRA problem	Solution	Solution type
1	EXTINGUISHER/TRUCK/CAMP	FIRE	MIXED
2	SURF/PADDLE/SKATE	BOARD	BACK
3	FOOT/BASKET/VOLLEY	BALL	BACK
4	JAY/MOCKING/FLU	BIRD	MIXED
5	DEW/COMB/BEE	HONEY	FRONT
6	LOSER/THROAT/SPOT	SORE	FRONT
7	DUCK/FOLD/DOLLAR	BILL	MIXED
8	BEAT/RATE/DISEASE	HEART	FRONT
9	LIGHT/BIRTHDAY/STICK	CANDLE	MIXED
10	OPERA/HAND/DISH	SOAP	MIXED
11	PALM/SHOE/HOUSE	TREE	MIXED
12	PLANE/SHIP/LINE	AIR	FRONT
13	MOUSE/DEATH/SAND	TRAP	BACK
14	RING/BOOK/SILK	WORM	BACK
15	SHINE/LIGHT/BEAM	SUN/MOON	FRONT
16	SENSE/COURTESY/PLACE	COMMON	FRONT
17	WET/DRY/FARM	LAND	BACK

Item number	CRA problem	Solution	Solution type
18	SPOON/CLOTH/CARD	TABLE/TEA	MIXED
19	RAIN/TEST/STOMACH	ACID	MIXED
20	COVER/ARM/WEAR	UNDER/BAND	FRONT
21	CUT/CREAM/WAR	COLD	FRONT
22	BABY/SPRING/CAP	SHOWER	MIXED
23	OFF/MILITARY/FIRST	BASE	BACK
24	LINE/TABLE/SCALE	TIME	FRONT

Note. Solution type refers to whether the solution word must be appended to the front of the cue words, to the back, or both ("mixed").

B. Data preprocessing

B.1. CRA Task

To account for typing or orthographic mistakes in the CRA responses, we computed the Levenshtein distance between each response and solution strings, which measures the minimum number of single-character edits required to transform one string into another. For instance, the two strings “RRAO” and “TRAP” have a Levenshtein distance of 2. If the computed distance was less than or equal to 2, the response was manually coded as correct when it matched the correct response. Furthermore, in contrast to previous studies, we included alternative correct answers for three CRA items when participants’ responses formed meaningful compound words. When both experimenters agreed that an “incorrect” response was valid, it was accepted as correct and added to the list of possible solutions for that CRA item (see Appendix F).

B.2. Animal Fluency Task

To preprocess the animal fluency data for our associative thinking analyses, we used the *SemNetCleaner* package in R (Christensen, 2019), following Christensen and Kenett (2021). The *SemNetCleaner* package offers a standardized method for preprocessing raw verbal fluency data. First, the pipeline removes non-category members (e.g., tree, unicorn) and duplicate participant responses. Subsequently, it corrects spelling errors, compound responses, root word variations, and continuous strings automatically. Finally, one of the researchers manually corrected the remaining words that were not recognized by the software.

C. Instructions

C.1. Two-Response Paradigm

After signing a consent form, participants received the following instructions:

Please read these instructions carefully!

In this experiment you will have to solve 24 word puzzle problems and 9 practice problems.

The word problems will be presented to you one after the other and you should not pause between them. You can take a short break in the middle of the experiment.

It is important that you complete the experiment in one sitting and without distractions.

During the experiment, in each trial you will be presented with three words. The goal is to **find a fourth word that you can attach to each of these three words so that three new meaningful compound words are created**.

For example, the three words BRUSH/PASTE/PICK are connected by the word TOOTH, because with the word TOOTH the compound words TOOTHBRUSH/TOOTHPASTE/TOOTHPICK can be formed.

Similarly, the three words RIVER/NOTE/ACCOUNT are connected by the word BANK, because with the word BANK the compound words RIVERBANK/BANKNOTE/BANK ACCOUNT can be formed.

For every word puzzle, the solution is always a word that you can only add either to the front or to the back of each of the three words.

Once you have found the solution, enter your answer and press “Enter”.

If you really can’t find the solution after thinking about it, press “Enter” to go to the next puzzle.

Do your best to solve as many puzzles as possible.

After you have solved a word puzzle, indicate **your confidence in your solution**. You can do this by using the cursor of the mouse to choose a position on a horizontal scale between “low confidence” and “high confidence”.

Finally, you must indicate **whether you have solved this word puzzle “with Aha!” or “without Aha!”**.

With Aha!: with an Aha! experience you become aware of the solution suddenly and clearly. This can be accompanied by a sense of revelation and relief.

Without Aha!: Unlike an Aha! feeling, finding a solution with analysis is characterized by a step-by-step search process.

Imagine a dark room that is suddenly lit up (with Aha!) or slowly lit with a dimmer switch (without Aha!). We ask you to indicate after each word puzzle if you have solved it "with Aha!" or "without Aha!".

We are going to start with 3 practice puzzles.

For each puzzle, a fixation cross will appear first. Then, the three words will be presented.

Participants were then given three CRA practice trials with no cognitive load or deadline. Importantly, they received feedback about the correct response on each trial. They were then familiarized with the two-response paradigm:

That was the first practice.

In this experiment, we want to know what **your initial, intuitive response** to these word puzzles is and **how you respond after you have thought about the word puzzles for some more time**.

Hence, as soon as the word puzzle is presented, we will ask you to enter your **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.

Next, the word puzzle 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 enter your **final response**. You will have as much time as you need to indicate your second response.

In sum, keep in mind that it is really crucial that **you give your first, initial response as fast as possible**.

Afterwards, you can take **as much time as you want to reflect on the word puzzle and select your final response**.

> We are going to start with a couple of practice problems. From now on we will not be providing feedback on your responses to the practice problems.
>
> First, a fixation cross will appear.
>
> Then, the word puzzle will be presented.
>
> As we told you, we are interested in your initial, intuitive response. 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 **8 seconds** (Experiment 1) / **6 seconds** (Experiment 2).
>
> When there are 2 seconds (Experiment 1) / 1 second (Experiment 2) left, the background color will turn to yellow to let you know that the deadline is approaching. Please make sure to **enter a response before the deadline passes**. If you really can't think of a word, make sure to press "Enter" instead of letting the trial pass.
>
> Next, **the word puzzle 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 enter your **final response.**
>
> After you have made your choice and click on it, you will be automatically taken to the next page.

Participants were then given two two-response trials without the concurrent cognitive load, without feedback this time. They were then introduced to the cognitive load task:

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

Participants had then to do two practice trials for the cognitive load task without the CRA problems. Following this, they were provided with the following instructions:

> In the actual study you will need to memorize the pattern while you give your initial, intuitive response to the word puzzle. The pattern is briefly presented before each problem. You do not have to memorize a pattern during your second, final response.

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 word puzzle. The most important thing is to correctly memorize the pattern.

First, **try to concentrate on the memorization task**, and then try to solve the word puzzle.

As a next step, you can practice this with two word puzzles.

After those two last practice trials, participants received the following instructions:

This is the end of practice.

The questions will be presented to you one after the other and you should not pause between them. After the first 12 questions, you will be invited to take a short break.

Please **stay as focused as possible** and try to solve as many problems as you can.

Remember, if you really don't know the response to a word puzzle, make sure to press "Enter" instead of letting the trial pass.

C.2. Animal Fluency Task (Experiment 2)

At the end of the CRA problems in Experiment 2, participants received the following instructions:

In this task you have to **write the names of as many different animals as you can think of in three minutes**, as quickly as possible.

Type the name of an animal and press **Enter** to type the next one.

You will be automatically taken to the next page when the three minutes have passed. Please keep writing animal names until the time runs out.

D. The "Aha" Experience Indexes Response Fluency Rather Than Sound Intuiting

The high prevalence of sound intuiting we observed could potentially be driven by insight problem-solving (i.e., the "Aha!" moment). Indeed, if insight problem-solving relies on intuitive, Type 1 processes to reach the correct solution, a high prevalence of insight responses may account for these findings. At the phenomenological level, insight problem-solving was associated with higher confidence ratings in both the initial and final response stages. Fig. A.1a provides a summary of the initial and final confidence levels as a function of solution type.

To analyze these findings, we built a beta generalized mixed-effects model for each experiment, analyzing confidence as a function of response stage, insight, and their interaction using sum coding. The results from Experiment 1 indicated that confidence ratings were significantly higher in the final than in the initial response stage, $OR = 1.35$, $p < .001$, Cohen's $d = 0.17$, 95% CI [0.13, 0.20], as well as for insight compared to non-insight solutions, $OR = 1.99$, $p < .001$, Cohen's $d = 0.38$, 95% CI [0.32, 0.44]. The interaction term between insight and response stage was also significant, $OR = 0.90$, $p < .001$, Cohen's $d = -0.06$, 95% CI [-0.08, -0.04]. Post-hoc tests showed that the effect of response stage on confidence was significant both for non-insight responses, $OR = 0.44$, $Z = -11.52$, $p < .001$, and for insight responses, $OR = 0.68$, $Z = -5.129$, $p < .001$. The results were similar for Experiment 2, with confidence ratings significantly higher in the final compared to the initial response stage, $OR = 1.30$, Cohen's $d = 0.14$, 95% CI [0.12, 0.17]. Insight solutions also showed higher confidence compared to non-insight solutions, $OR = 1.86$, Cohen's $d = 0.34$, 95% CI [0.29, 0.40]. The interaction between insight and response stage was again significant, $OR = 0.92$, Cohen's $d = -0.05$, 95% CI [-0.07, -0.02]. Post-hoc tests showed that the effect of response stage on confidence was significant both for non-insight responses, $OR = 0.50$, $Z = -10.47$, $p < .001$, and for insight responses, $OR = 0.70$, $Z = -4.64$, $p < .001$. Overall, the

confidence results indicate that the effect of response stage was smaller for insight responses, as participants already had high confidence in the initial response stage for insight responses.

In addition, insight problem-solving was also associated with shorter response times, both in the initial (Experiment 1: *M* insight = 4.4 s vs. *M* non-insight = 5.5 s; Experiment 2: *M* insight = 3.6 s vs. *M* non-insight = 4.3 s) and the final response stage (Experiment 1: *M* insight = 9.3 s vs. *M* non-insight = 22.5 s; Experiment 2: *M* insight = 7.6 s vs. *M* non-insight = 20 s). Fig. A.1b summarizes those findings.

To test these results statistically, we used a log-linear mixed-effects model on reaction times as a function of insight, response stage, and their interaction using sum coding, separately for Experiment 1 and Experiment 2. Results from Experiment 1 showed that participants were significantly slower in the final than in the initial response stage, exp(β) = 1.26, $t(3275.26) = 20.19$, $p < .001$, Cohen's $d = 0.81$, 95% CI [0.73, 0.89], and significantly faster for insight compared to non-insight responses, exp(β) = 0.85, $t(101.32) = -8.32$, $p < .001$, Cohen's $d = -0.54$, 95% CI [-0.69, -0.39]. The interaction term between insight and response stage was also significant, indicating that the effect of response stage was smaller for insight responses, exp(β) = 0.85, $t(3259.97) = -14.64$, $p < .001$, Cohen's $d = -1.05$, 95% CI [-1.19, -0.9]. Post-hoc tests showed that the effect of response stage on reaction times was significant both for non-insight responses, $t(3268) = -26.13$, $p < .001$, and for insight responses, $t(3263) = -3.61$, $p < .001$. The results were similar for Experiment 2, with participants significantly slower in the final compared to the initial response stage, exp(β) = 1.29, $t(2499.43) = 19.95$, $p < .001$, Cohen's $d = 0.81$, 95% CI [0.73, 0.90], and significantly faster for insight compared to non-insight responses, exp(β) = 0.84, $t(101.78) = -7.25$, $p < .001$, Cohen's $d = 0.54$, 95% CI [0.39, 0.70]. The interaction between insight and response stage was also significant, exp(β) = 0.82, $t(2491.69) = -15.28$, $p < .001$, Cohen's $d = -1.25$, 95% CI

[-1.42, -1.09]. Post-hoc tests showed that the effect of response stage on reaction times was significant both for non-insight responses, $t(2474) = -27.18$, $p < .001$, and for insight responses, $t(2489) = -3.00$, $p = .0027$. Insight problem-solving was thus associated with higher confidence and faster reaction times in both experiments, consistent with the possibility that insight may account for sound intuition in our results.

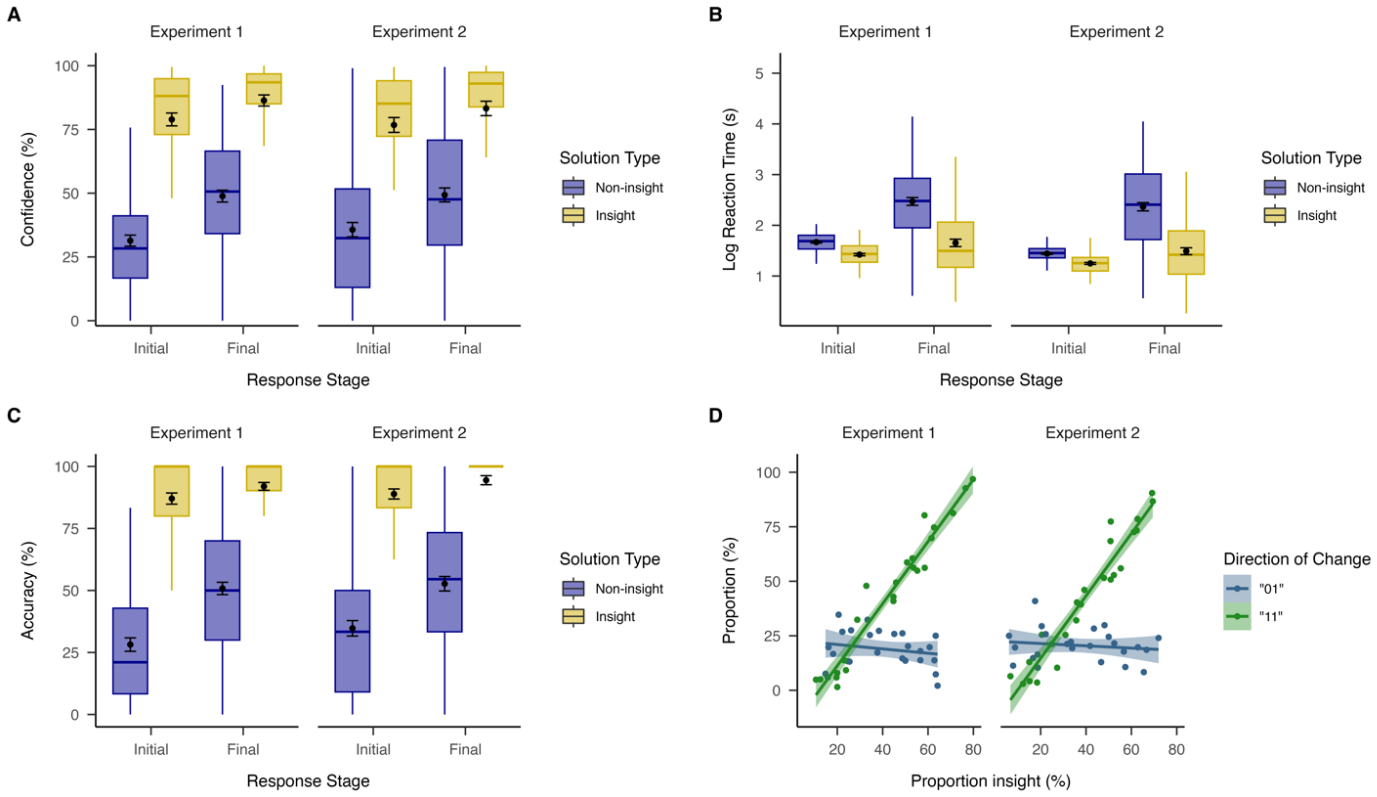

Figure A.1. Behavioral correlates of insight in Experiments 1 and 2. **a)** Response confidence as a function of response stage and solution type. **b)** Logged reaction time as a function of response stage and solution type. **c)** Accuracy as a function of response stage and solution type. **d)** By-item proportion of "11" and "01" responses as a function of the proportion of insight in the initial response stage ("11" responses) and in the final response stage ("01" responses). In the boxplots, the lower and upper hinges 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.

A central question revolves around whether insight solutions are affected by the initial stage of the paradigm compared to non-insight solutions. Indeed, if insight does not rely on Type 2, deliberate processes, insight responses should not (or less) be affected by the constraints of the initial response stage where deliberation is minimized. Fig. A.1c provides a summary of the initial and final accuracy as a function of solution type. However, at the item level, the proportion of reported initial insight responses was almost perfectly correlated with the proportion of "11" responses in Experiment 1, $r(22) = .97$, $p < .001$, 95% CI [.94, .99], as well as in Experiment 2, $r(22) = .96$, $p < .001$, 95% CI [.92, .98]. The proportion of final insight did not correlate with the proportion of "01" responses in Experiment 1, $r(22) = -.14$, $p = .50$, 95% CI [-.52, .28], or in Experiment 2, $r(22) = -.14$, $p = .50$, 95% CI [-.52, .28]. This indicates that the simpler a problem was from the initial stage, the more participants tended to report solving it through insight, while solving a problem after an initial incorrect response (or no response at all) was not associated with insight (Fig. A.1d). Thus, any analysis testing the relationship between insight problem-solving and sound intuiting should take this item-level confound into account.

To test the theoretical possibility that problems solved through insight would be less affected by the initial response stage, we built a mixed-effect logistic model on accuracy using insight, response stage, and their interaction as fixed effects using sum coding. We also added the final difficulty of the item as a control variable in the model since insight is positively correlated with item difficulty. In Experiment 1, there was a significant effect of insight on accuracy, $OR = 4.79$, $p < .001$, Cohen's d = 0.86, 95% CI [0.71, 1.05], response stage, $OR = 2.00$, $p < .001$, Cohen's d = 0.38, 95% CI [0.29, 0.48], as well as item difficulty, $OR = 1.07$, $p < .001$, Cohen's d = 0.04, 95% CI [0.03, 0.04]. However, the interaction between insight and response stage did not reach significance, $OR = 0.90$, $p = .19$, Cohen's $d = -0.06$, 95% CI [-0.14, 0.04]. In Experiment 2, the

results were identical. There was a significant effect of insight on accuracy, $OR = 4.29$, $p < .001$, Cohen's $d = 0.80$, 95% CI [0.71, 0.91], response stage, $OR = 2.22$, $p < .001$, Cohen's $d = 0.44$, 95% CI [0.35, 0.55], as well as item difficulty, $OR = 1.06$, $p < .001$, Cohen's $d = 0.03$, 95% CI [0.03, 0.03]. Once again, the interaction between insight and response stage did not reach significance, $OR = 1.04$, $p = .68$, Cohen's $d = 0.02$, 95% CI [-0.07, 0.12]. Thus, although insight was consistently associated with higher accuracy, we did not find evidence supporting a distinct impact of the cognitive load and response deadline on insight solutions compared to non-insight solutions.

Overall, these findings suggest that insight problem-solving does not depend on distinct, intuitive cognitive mechanisms that could drive sound intuition. Rather, insight seems to index higher response fluency, characterized by easier problems, higher confidence, and faster responses.

E. Mixed-Effects Models in the CRA task

In the CRA, we used (generalized) linear mixed models, allowing trial-by-trial analysis while accounting for variation across both participants and items. Unlike traditional ANOVA, mixed models provide greater flexibility by avoiding the need for prior averaging, minimizing Type I error, managing unbalanced datasets, and improving predictive precision and generalization through partial pooling, where data for individual participants or items is informed by the entire dataset (Baayen et al., 2008).

To analyze accuracy and direction of change, we used binomial generalized mixed models (Bolker et al., 2009). For the supplementary models on response confidence, we used mixed-effects beta regression (Verkuilen & Smithson, 2012), with adjustments to keep data within a .005 to .995 range before back-transforming it to its original scale (Smithson & Verkuilen, 2006). For reaction times analyses, the trials could be censored at the deadline in the initial response stages of Experiment 1 and 2. To account for this, we used censored (Tobit) log-normal mixed-effects models (Bürkner, 2017), treating trials hitting the deadline as right-censored rather than excluding them. For the exploratory insight reaction time analyses, insight ratings were only available when participants provided a valid response, so trials over the response deadline in the initial response stage were excluded by definition. These analyses were therefore conducted with conventional log-transformed linear mixed models, with results back-transformed to seconds for interpretation.

To determine the optimal random structure for each analysis, we first identified the maximal model supported by the data, followed by backward stepwise elimination using the likelihood ratio test to maximize statistical power (Matuschek et al., 2017). For the binomial models, we assessed the significance of fixed effects using parametric bootstrapping with 1000 iterations (Booth, 1995). For reaction times, we evaluated fixed-effect significance with the Kenward-Roger t-test (Kenward

& Roger, 1997). Censored reaction times models were estimated in a Bayesian framework, due to convergence issues in the frequentist framework. We used weakly informative priors using four chains of 4,000 iterations each (2,000 warm-up), yielding 8,000 post–warm-up draws in total. Posterior estimates are reported with 95% credible intervals (*CrI*) and posterior probabilities of direction (*pd*). In the case of the supplementary mixed-effects beta regression models on confidence, we used Wald Z-tests to assess significance due to convergence issues encountered during the bootstrapping procedure.

In models with interaction terms, we applied sum coding as our contrast scheme. Significant interactions were further explored via post-hoc tests on estimated marginal means, with Holm-Bonferroni correction for multiple comparisons. For generalized mixed models, effect sizes are reported by converting odds ratios to Cohen's *d* (Borenstein et al., 2009).

We used the following R packages for the mixed-effects models analyses: *bayestestR* (Makowski et al., 2019), *brms* (Bürckner, 2017), *buildmer* (Voeten, 2020), *emmeans* (Lenth et al., 2019), *glmmTMB* (Brooks et al., 2017), *lmerTest* (Kuznetsova et al., 2017), *lme4* (Bates et al., 2014) and *parameters* (Lüdecke et al., 2020).

F. Semantic Network Modelling

F.1. Semantic Networks Construction

To estimate the semantic networks based on the animal fluency task data, we used three out of the four methods available in the *SemNet* package in R (Christensen & Kenett, 2023): the Correlation-Based Network (CbN) method, the Pathfinder Network (PN) method and the Naive Random Walk (NRW) method. The Community Network method was excluded from our analyses, as the estimated networks did not significantly differ from randomly generated networks (see below). The networks were estimated separately for the two groups based on the median split of the proportion of "11" responses (i.e., "High 11" vs. "Low 11" group).

In general, we used the default *SemNet* parameters but adjusted them when the estimated networks did not significantly differ from randomly generated ones. Below, we outline the methods and parameters used for network estimation (for a detailed explanation of these methods, see Zemla & Austerweil, 2018):

1) The CbN method builds a semantic network by analyzing co-occurrences within a binary response matrix, where rows represent individual participants and columns represent unique animal names given across participants in the verbal fluency task. Each cell in the matrix is filled with a "1" if the corresponding participant mentioned the animal and a "0" otherwise. Responses not provided by at least three participants per group were excluded to control for confounding factors such as the number of nodes and edges across groups (Borodkin et al., 2016). Additionally, we included only the responses given across both groups to ensure that our comparison was based solely on structural differences within the same nodes across the two networks. We then estimated the networks based on the co-

occurrence of responses across the group. Here, we used Pearson's pairwise correlation to compute the pairwise similarities between each column (e.g., Kenett et al., 2013), which resulted in an association matrix. The triangulated maximally filtered graph (TMFG; Christensen et al., 2018a) was subsequently applied to the association matrix to ensure that only the strongest, most relevant connections were maintained while every node remained connected in the network.

2) The PN method (Quirin et al., 2008) also creates a network from the binary response matrix. Here again, we only kept the responses that were given across both groups, ensuring that our comparison would be based on structural differences in the organization of the nodes between the two networks. The PN method uses a proximity measure such as Euclidean distance to build a proximity matrix. It then keeps the path which has the shortest distance between every pair of nodes to build a network that emphasizes minimal distances. The method is parameterized by the two parameters *q* and *r*, which govern the number of steps between nodes and the computation of distance, respectively. The *SemNet* package sets these two parameters to build the sparsest possible network (i.e., the one with the fewest number of edges) from the proximity matrix, following Zemla and Austerweil (2018).
3) The NRW method (Jun et al., 2015) assumes that the fluency lists generated by the participants arise from an uncensored random walk, "stepping" from one node to another in the semantic network. The NRW method infers an edge between each adjacent pair of responses in the participants' fluency lists, thereby assuming that adjacent responses have a higher likelihood of being interrelated. To minimize the number of spurious connections in the network, we applied a threshold to remove pairs that appeared less than three times across participants to have an edge in the network (Lerner et al., 2009). In the NRW method, because response order is important, removing responses from participants' fluency lists to

control for the differing number of responses between groups is not possible. However, the difference in verbal fluency between the "High 11" group (M = 36.9 words) and the "Low 11" group (M = 39.4 words) unexpectedly favored the "Low 11" group, was minimal and statistically non-significant, $t(94.64) = 1.07$, $p = .28$. Verbal fluency should thus not bias our NRW findings–and even less so in the expected direction of our predictions.

F.2. Comparison to Random Networks

To ensure that the generated networks had different structures from random networks with the same number of edges, nodes, and degree sequence (i.e., connections per node), we compared the network metrics (i.e., ASPL, CC, and Q) of our networks against those of randomly generated networks. For each network estimation method and response group (i.e., "Low 11" and "High 11"), we generated 1000 random networks to create a sampling distribution of the network metrics to compute a p-value for the original group's network measures (Kenett et al., 2013). A significant result indicates that the network's structure differs from that of an equivalent random network for this particular network metric.

The results of this analysis revealed that the generated networks significantly differed from randomly generated networks for both groups across all network estimation methods and metrics (all $p < .001$). However, this was not the case for the Community Network method on the Q and ASPL metrics, which led us to exclude this method from our analyses.

F.3. Case-wise Bootstrap Analysis

Following Luchini et al. (2023), we used case-wise bootstrap analysis (Efron, 1979) to assess the difference in the structure of the semantic memory network between the "High 11" and the "Low 11" groups. Since the group-level estimation of network metrics provides only a single

value per group, the bootstrapping approach allows us to test whether any differences between the two groups are significant. We used the with-replacement bootstrapping of the *SemNet* package with 1000 iterations to compute the three network metrics (i.e., ASPL, CC, and Q) for each resampled group's network. Independent-samples t-tests were then conducted for each network metric to assess whether the differences between the two groups were significant. The results are presented in the tables A.2-4.

Table A.2. Independent *t*-tests results comparing network metrics between the "Low 11" and the "High 11" groups for the Correlation-Based Network method.

Variable	Mean Low 11 (SD)	Mean High 11 (SD)	95% CI	t(1998)	p-value	d	Direction
ASPL	4.405 (0.401)	4.346 (0.41)	[0.023, 0.094]	3.23	**.001**	0.144	Low 11 > High 11
CC	0.704 (0.007)	0.707 (0.007)	[-0.003, -0.002]	-7.58	**< .001**	0.339	Low 11 < High 11
Q	0.709 (0.013)	0.709 (0.013)	[-0.001, 0.001]	-0.23	.82	0.01	Low 11 = High 11

Note. ASPL: Average Shortest Path Length; CC: Clustering Coefficient; Q: Modularity; CI = Confidence Interval. A higher ASPL indicates a more spread-out network structure, a higher CC indicates greater interconnectivity within the network, and a higher Q indicates stronger modularity. *p*-values less than .05 are shown in bold.

Table A.3. Independent *t*-tests results comparing network metrics between the "Low 11" and the "High 11" groups for the Pathfinder Network method.

Variable	Mean Low 11 (SD)	Mean High 11 (SD)	95% CI	t(1998)	p-value	d	Direction
ASPL	3.674 (0.212)	3.553 (0.22)	[0.102, 0.14]	12.53	**< .001**	0.56	Low 11 > High 11
CC	0.346 (0.03)	0.387 (0.035)	[-0.043, -0.038]	-27.84	**< .001**	1.25	Low 11 < High 11
Q	0.52 (0.046)	0.454 (0.058)	[0.062, 0.071]	28.63	**< .001**	1.28	Low 11 > High 11

Note. ASPL: Average Shortest Path Length; CC: Clustering Coefficient; Q: Modularity; CI = Confidence Interval. A higher ASPL indicates a more spread-out network structure, a higher CC indicates greater interconnectivity within the network, and a higher Q indicates stronger modularity. *p*-values less than .05 are shown in bold.

Table A.4. Independent *t*-tests results comparing network metrics between the "Low 11" and the "High 11" groups for the Naïve Random Walk method.

Variable	Mean Low 11 (SD)	Mean High 11 (SD)	95% CI	t(1998)	p-value	d	Direction
ASPL	3.624 (0.24)	3.606 (0.446)	[-0.013, 0.05]	1.13	.26	0.05	Low 11 = High 11
CC	0.084 (0.014)	0.086 (0.014)	[-0.003, -0.001]	-3.04	**.002**	0.14	Low 11 < High 11
Q	0.459 (0.026)	0.439 (0.03)	[0.018, 0.023]	16.47	**< .001**	0.74	Low 11 > High 11

Note. ASPL: Average Shortest Path Length; CC: Clustering Coefficient; Q: Modularity; CI = Confidence Interval. A higher ASPL indicates a more spread-out network structure, a higher CC indicates greater interconnectivity within the network, and a higher Q indicates stronger modularity. *p*-values less than .05 are shown in bold.

F.4. Effect Sizes Pooling

As shown in Tables A.2-4, the results of our semantic network analyses did not always converge. Specifically, the CbN method result for the Q metric was non-significant, as was the NRW method result for the ASPL metric. We thus computed pooled standardized effect sizes, accounting for the dependence between these measures (Borenstein et al., 2021), as reported in Table A.5. We corrected for the fact that the effect sizes are likely correlated across methods by using the *aggregate* function in the *metafor* package in R (Viechtbauer, 2010). Given that the degree of correlation between our network methods is unknown, we assumed a conservative correlation of $\rho = .8$ for sampling errors within clusters. Importantly, sensitivity analyses across the full range of possible correlation values did not change the results in terms of significance.

Table A.5. Aggregated semantic memory network metrics comparison between the "Low 11" and the "High 11" groups.

Parameter	Hedge's *g*	95% CI	*Z*	*p*	Direction
ASPL	-0.25	[-0.33, -0.17]	-6.01	**< .001**	Low 11 > High 11
CC	0.61	[0.52, 0.7]	13.62	**< .001**	Low 11 < High 11
Q	-0.66	[-0.75, -0.57]	-14.61	**< .001**	Low 11 > High 11

Note. ASPL = Average Shortest Path Length; CC = Clustering Coefficient; Q = Modularity; CI = Confidence Interval. A higher ASPL indicates a more spread-out network structure, a higher CC indicates greater interconnectivity within the network, and a higher Q indicates stronger modularity. A positive effect size indicates that the specific component was larger in the High "11" group network, while a negative effect size indicates it was smaller. Effect sizes were aggregated assuming an inter-method correlation of $\rho = .8$. *p*-values less than .05 are shown in bold.

G. Estimated Models

In this section, we report the models presented in the results section of the paper. Although we did not preregister specific analyses, our preregistration indicated that we would analyze accuracy and direction of change based on item difficulty (three levels: easy, medium, and hard). However, we eventually chose to use semantic similarity between cue words and the solution word in each CRA item as a continuous proxy for item difficulty instead, given its theoretical relevance.

Table A.6. Binomial generalized mixed-effects model on accuracy as a function of response stage and experiment, using sum coding.

accuracy ~ 1 + response_stage + experiment + response_stage:experiment

+ (1 + response_stage | subject) + (1 | item_number)

Predictors	OR	95% CI	p.value
Intercept	1.32	[0.65, 2.71]	.43
Response stage	1.89	[1.74, 2.04]	**< .001**
Experiment	0.98	[0.85, 1.12]	.44
Response stage:Experiment	1.05	[0.98, 1.13]	.17

Random effects

Group	Parameter	SD
Subject	(Intercept)	0.86
Item	(Intercept)	1.79
Subject	Response stage	0.24
Subject	Cor (Intercept x Response stage)	0.18

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 A.7. Binomial generalized mixed-effects model on the probability of giving a "11" response (vs. a "01" response) contrasting Experiment 1 and Experiment 2, using dummy coding.

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

Predictors	OR	95% CI	p.value
Intercept (Experiment 1)	1.98	[1.13, 3.54]	**.014**
Experiment 2	0.87	[0.62, 1.21]	.36

Random effects

Group	Parameter	SD
Subject	(Intercept)	0.90
Item	(Intercept)	1.24

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.

Table A.8. Linear regression model output for the by-subject proportion of "11" responses as a function of forward flow, education level and verbal fluency in Experiment 2.

Predictors	Estimates	95% CI	p.value
(Intercept)	40.60	[18.74, 62.47]	**< .001**
Forward Flow	6.01	[0.92, 11.09]	**.021**
Education Level	1.12	[-3.98, 6.23]	.66
Verbal Fluency	0.13	[-0.29, 0.55]	.54
Observations:	98		
R^2 / R^2 adjusted:	0.10 / 0.07		

H. Confidence Analyses

H.1. Intuitive Error Sensitivity

Are participants sensitive to their errors in the initial response stage of the paradigm? In Experiment 1, participants reported higher confidence for correct responses (M = 77.5, SD = 21.1) compared to incorrect responses (M = 21.6, SD = 17.4) in the initial stage. A beta regression mixed-effects model on the initial confidence as a function of accuracy using dummy coding confirmed that the initial confidence ratings on correctly solved CRA puzzles were significantly higher than those for incorrect responses, OR = 10.26, $p < .001$, Cohen's d = 1.28, 95% CI [1.10, 1.46]. The results were similar in Experiment 2, where participants also reported a higher confidence for correct responses (M = 77.2, SD = 26.1) compared to incorrect responses (M = 24.2, SD = 20), OR = 8.09, $p < .001$, Cohen's d = 1.15, 95% CI [0.98, 1.33]. Participants were thus able to recognize that they had not converged on the correct solution in the CRA, even when deliberation was minimized.

H.2. Initial Confidence and Response Change

Another question concerns whether lower confidence in the initial response stage predicts higher accuracy in the final response stage. To test this, we examined initial confidence ratings as a function of the direction of change. Crucially, we focused on initially incorrect responses (i.e., "00" or "01" responses) to control for initial accuracy. If lower confidence in the initial stage leads to higher final accuracy, we should observe lower confidence for "01" responses compared to "00" responses.

The observed confidence differences were small in both experiments in Experiment 1 ("00" responses: M = 22.3, SD = 18.9; "01" responses: M = 20.9, SD = 20.4) and in Experiment 2 ("00"

responses: $M = 22.6$, $SD = 19$; "01" responses: $M = 23.8$, $SD = 23.2$). To test this statistically, we built two mixed-effects logistic regression models to predict final accuracy as a function of initial confidence, subsetting on incorrect initial responses (e.g., "00" or "01" responses). In Experiment 1, initial confidence did not significantly predict final accuracy, $OR = 0.93$, $p = .52$, Cohen's d = -0.04, 95% CI [-0.17, 0.08]. The pattern was the same in Experiment 2, $OR = 0.94$, $p = .62$, Cohen's d = -0.04, 95% CI [-0.18, 0.11]. Thus, in both experiments, initial confidence did not predict final accuracy.

I. Comparison of Item Difficulty Across Studies

A potential criticism of our work could be that we only observed sound intuiting due to the selection of easier items than usually found in the literature. To address this concern, we can directly compare our results to other studies using the CRA. We computed the combined average accuracy in the final response stage across both of our studies. The outcome (61%) was higher than the accuracy reported in Kounios et al. (2006) (46.2%) and Chein and Weisberg (2014) (42%), similar to the accuracy reported in Cranford and Moss (2012) (55%), and lower than the accuracy reported in Stuyck et al. (2021) (78%) or Stuyck et al. (2022) (No-Load condition; 79,9%). In addition, we meticulously checked responses for spelling accuracy before coding them as correct or incorrect, and we acknowledged alternative responses as valid solutions to problems when they were correct (see Appendix H). This may have further boosted the accuracy of our studies.

J. Impact of Recoded Missed Deadlines on Intuitive Performance

It is possible that participants safeguarded their performance on the CRA over the deadline or the cognitive load task in the initial response stage. Since we excluded trials where participants failed to answer before the deadline or to give a correct answer to the cognitive load task, such a response strategy coupled with the trials' exclusion could lead to an overestimation of correct intuitive responses.

To investigate such a possibility, we correlated the by-item proportion of missed deadlines and loads with item difficulty (computed as 100 minus the mean accuracy in the final response stage Results showed that indeed the item difficulty correlated strongly with the proportion of missed deadlines, $r(22) = .84$, $p < .001$, 95% CI [.66, .93] in Experiment 1 and $r(22) = .78$, $p < .001$, 95% CI [.56, .90] in Experiment 2. However, item difficulty did not correlate significantly with the proportion of failed cognitive loads, $r(22) = .19$, $p = .37$, 95% CI [-.23, 0.55] in Experiment 1 and $r(22) = .08$, $p = .70$, 95% CI [-.33, 0.47] in Experiment 2. This result indicates that participants missed more deadlines when the item was harder, possibly suggesting that participants safeguarded their performance on the CRA at the expense of the time constraint.

We thus performed an additional analysis where we recomputed the mean non-correction rate while conservatively coding the missed deadline trials as incorrect responses. The average non-correction rate was slightly lower in this conservative analysis: $M = 70.4\%$ ($SD = 17.8$) in Experiment 1 and $M = 63.6\%$ ($SD = 21.1$) in Experiment 2.

To statistically assess whether there was a higher occurrence of sound intuiting ("11") as opposed to correct deliberate responses following an initial incorrect response ("01"), we focused exclusively on final correct responses, excluding "00" and "10" trials. We then built a mixed-

effects logistic regression model, which included random intercepts for both participants and items, with the experiment as a fixed effect. We used a dummy variable to code whether a trial was a "01" response (coded as 0) or a "11" response (coded as 1). Importantly, the estimated non-correction rates from the model were significantly different from 0 both for Experiment 1, M = 64%, 95% CI [51, 75] and Experiment 2, M = 56%, 95% CI [42, 68]. The effect of experiment was significant, OR = 0.71, p = .018, Cohen's d = -0.19, 95% CI [-0.36, -0.04], indicating that there was a small but significant decrease in the non-correction rate of Experiment 2 compared to Experiment 1. Thus, despite this slight decrease, our findings show that when participants gave a correct answer in the final response stage of the paradigm, they had already given a correct answer in the initial response stage most of the time. This finding confirms the robustness of our results, even when recoding the missed deadlines trials as incorrect.

Appendix References

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

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

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

Booth, J. (1995). Bootstrap methods for generalized linear mixed models with applications to small area estimation. In *Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modelling Innsbruck, Austria, 10–14 July, 1995* (pp. 43-51). Springer New York. https://doi.org/10.1007/978-1-4612-0789-4_6

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Converting among effect sizes. *Introduction to meta-analysis*, *147*(4), 45-49.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). Multiple outcomes or time-points within a study. In M. Borenstein, L. V. Hedges, J. P. Higgins, & H. R. Rothstein (Eds.), *Introduction to meta-analysis* (pp. 225-238). John Wiley & Sons.

Borodkin, K., Kenett, Y. N., Faust, M., & Mashal, N. (2016). When pumpkin is closer to onion than to squash: The structure of the second language lexicon. *Cognition*, *156*, 60–70. https://doi.org/10. 1016/j.cognition.2016.07.014

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

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chein, J. M., & Weisberg, R. W. (2014). Working memory and insight in verbal problems: Analysis of compound remote associates. *Memory & Cognition*, *42*(1), 67–83. https://doi.org/10.3758/s13421-013-0343-4

Christensen, A. P. (2019). *SemNetCleaner: An automated cleaning tool for semantic and linguistic data.* R package version, 1(0).

Christensen, A. P., & Kenett, Y. N. (2023). Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *Psychological methods*, *28*(4), 860–879. https://doi.org/10.1037/met0000463

Christensen, A. P., Kenett, Y. N., Cotter, K. N., Beaty, R. E., & Silvia, P. J. (2018). Remotely close associations: Openness to experience and semantic memory structure. *European Journal of Personality*, *32*, 480–492. https://doi.org/10.1002/per.2157

Cranford, E. A., & Moss, J. (2012). Is insight always the same? A protocol analysis of insight in compound remote associate problems. *The Journal of Problem Solving*, 4(2), Article 8. https://doi.org/10.7771/1932-6246.1129

Efron, B. (1979). Bootstrap method: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Jun, K.-S., Zhu, X., Rogers, T., Yang, Z., & Yuan, M. (2015). Human memory search as initial-visit emitting random walk. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 1072–1080). Retrieved from https://proceedings.neurips.cc/ paper/2015/hash/dc6a70712a252123c40d2adba6a11d84-Abstract.html

Kenett, Y. N., Wechsler-Kashi, D., Kenett, D. Y., Schwartz, R. G., Ben Jacob, E., & Faust, M. (2013). Semantic organization in children with cochlear implants: Computational analysis of verbal fluency. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00543

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558

Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., & Jung-Beeman, M. (2006). The prepared mind: neural activity prior to problem presentation predicts subsequent solution by sudden insight. *Psychological science*, *17*(10), 882–890. https://doi.org/10.1111/j.1467-9280.2006.01798.x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

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

Lerner, A. J., Ogrocki, P. K., & Thomas, P. J. (2009). Network graph analysis of category fluency testing. *Cognitive and Behavioral Neurology*, *22*, 45–52. https://doi.org/10.1097/WNN.0b013e318192ccaf

Luchini, S., Kenett, Y. N., Zeitlen, D. C., Christensen, A. P., Ellis, D. M., Brewer, G. A., & Beaty, R. E. (2023). Convergent thinking and insight problem solving relate to semantic memory network structure. *Thinking Skills and Creativity*, *48*, 101277. https://doi.org/10.1016/j.tsc.2023.101277

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

Makowski D, Ben-Shachar M, Lüdecke D (2019). "bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework." *Journal of Open Source Software*, *4*(40), 1541. https://joss.theoj.org/papers/10.21105/joss.01541.

Quirin, A., Cordón, O., Guerrero-Bote, V. P., Vargas-Quesada, B., & Moya-Anegón, F. (2008). A quick MST-based algorithm to obtain Pathfinder networks (∞, n - 1). *Journal of the American Society for Information Science and Technology*, *59*, 1912–1924. https://doi.org/10.1002/asi.20904

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

Stuyck, H., Aben, B., Cleeremans, A., & Van den Bussche, E. (2021). The Aha! moment: Is insight a different form of problem solving?. *Consciousness and cognition*, *90*, 103055. https://doi.org/10.1016/j.concog.2020.103055

Stuyck, H., Cleeremans, A., & Van den Bussche, E. (2022). Aha! under pressure: The Aha! experience is not constrained by cognitive load. *Cognition*, *219*, 104946. https://doi.org/10.1016/j.cognition.2021.104946

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software*, *36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03

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

Zemla, J. C., & Austerweil, J. L. (2018). Estimating semantic networks of groups and individuals from fluency data. *Computational brain & behavior*, *1*(1), 36–58. https://doi.org/10.1007/s42113-018-0003-7