Thinking Skills and Creativity 46 (2022) 101114

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc

Insight with stumpers: Normative solution data for 25 stumpers and a fresh perspective on the accuracy effect[☆]

Wendy Ross[a,*], Frédéric Vallée-Tourangeau[b]

[a] London Metropolitan University, United Kingdom
[b] Kingston University, United Kingdom

ARTICLE INFO

Keywords:
Insight
Accuracy effect
Stumpers

ABSTRACT

When people solve a problem, they can do in one of two ways - analytically or through insight. There is robust evidence showing that a problem solved insightfully is more likely to be correct than one solved through analysis, the so-called accuracy or correctness effect in insight research. However, the nature of the insight problems in the laboratory means that it is often not easy to disentangle whether a participant *feels* correct or whether she actually *is* correct. We report data from two studies using stumpers as stimuli. Stumpers are a form of riddle in which it is possible to generate a plausible but incorrect answer. Alongside normative data for 25 stumpers, we also demonstrate that insight is linked to certainty in the answer rather than whether the answer is correct or not and that certainty (subjective correctness) is a stronger predictor of the feeling of insight than objective correctness. The findings support work into false insight and further add to the understanding of the phenomenology of 'aha' moments.

1. Introduction

There are times when a problem solution comes to the mind of the problem solver without her being aware of how she got to the answer. Such a moment is marked by suddenness and a feeling of certainty. In the words of a participant in Hill and Kemp's (2018) qualitative study, "it suddenly clicked and it made perfect sense" (p. 206). This is the phenomenon of insight. It is something that is commonly experienced and is endowed with an easily recognisable character, but the causes and mechanisms are opaque to the person experiencing it and pose a challenge for researchers. Insight variously refers to a class of problems, a cognitive process based on information restructuring and a phenomenological reaction to uncovering new knowledge. The importance of both the affective and cognitive dimensions invites a more granular consideration of the role of emotions in the genesis of a new idea.

Insight researchers often preface their work with a description of an embedded, idiosyncratic process drawn from everyday experience or retrospective anecdotal reports of creative breakthroughs as a means to situate and justify the examination of the phenomenon under laboratory conditions. Moving from such a process to one which can be studied in the lab requires stimuli which can elicit this feeling. While the feeling of insight is common, it is unpredictable and so needs to be artificially generated (Friedlander & Fine, 2018). The scaled translation of a phenomenon from its manifestation in the world to an operationalized laboratory event is not unique to research on insight, of course, but such an idiosyncratic process requires the careful engineering of both affective and

[☆] We thank Susan Cooper, Fiona Campbell, Tom Ormerod and two anonymous reviewers for helpful comments on previous versions of this manuscript and Margaret Webb for constructive discussions on both the content and the data analysis strategy.

[*] Corresponding author at: Psychology Department, London Metropolitan University, 166-220 Holloway Road, N7 8DB, United Kingdom.
E-mail address: w.ross@londonmet.ac.uk (W. Ross).

https://doi.org/10.1016/j.tsc.2022.101114
Received 15 December 2021; Received in revised form 11 August 2022; Accepted 20 August 2022
Available online 24 August 2022

cognitive dimensions.

1.1. The accuracy/correctness effect in insight problem solving

Initial research into insight problem solving approached it as a task-based phenomenon and the discovery of the correct answer to an insight problem was assumed to have necessarily involved insightful processes (Chronicle, MacGregor & Ormerod, 2004; Ormerod, MacGregor & Chronicle, 2002; Weisberg, 1995). Investigations into this process therefore traditionally contrasted performance on these problems with performance on analytic problems. An analytic problem is one where the operators to reach the goal state are well known, and while the participants cannot conjure up the correct answer immediately upon seeing the problem description (e.g., what is 567 times 876?), the methodical and iterative application of known operators yields an answer (viz., 496,692), without eliciting the signature phenomenology of an "aha" moment. In contrast, an insight problem is designed to encourage an incorrect interpretation of the problem that informs an unproductive search for operators that might or might not be applicable to the problem (e.g., how do you throw a ping pong ball in such a way that it comes to a complete stop and without hitting anything reverses direction?). However, as evidence has accumulated that the same problem can be solved with or without a feeling of insight (Bowden & Jung-Beeman, 2003; Danek & Wiley, 2017; Webb, Little & Cropper, 2016), the question arises over which is the most likely marker of a correct answer – in other words, which is the "best" way to solve such a problem.

Danek and Salvi (2020) have proposed that a problem solution accompanied with a feeling of insight is more likely to be correct than one which does not yield that feeling. They draw on a considerable amount of data to support the proposal that the phenomenology of insight is a marker of solution accuracy. The majority of this work comes from research which contrasts analytical problem solving with solving a problem through insight and suggests that solutions generated through insight are more likely to be correct. Participants are invited to report whether they worked effortfully towards the solution or whether the answer came through insight. Across a range of tasks, when participants reported coming across the solution through insight, they were more likely to be correct than when they reported effortful processes (Kizilirmak, Gallisch, Schott & Folta-Schoofs, 2021; Laukkonen, Ingledew, Grimmer, Schooler & Tangen, 2021; Spiridonov, Loginov & Ardislamov, 2021; Stuyck, Aben, Cleeremans & Van den Bussche, 2021; Threadgold, Marsh & Ball, 2018). This can be measured by comparing whether experiencing "aha" makes it more likely to have a correct answer (the "accuracy effect") or whether average "aha" ratings are higher for correct than incorrect solutions (the "correctness effect"; Threadgold et al., 2018). Strickland, Wiley and Ohlsson (2022) collate a range of studies and show effect sizes ranging from 0.56 to 2.63 suggesting a robust effect (although it is worth noting that they were not able to replicate the accuracy effect for their work on visuo-spatial puzzles).

Danek and Salvi argue that the feeling of insight is a reflection of an objectively correct answer because it is only this which would yield a good "gestalt". The feeling of "aha", therefore, is rooted in the qualitative differences between correct and incorrect answer solutions; the key assumption here is that only a correct answer emerges whole as a good gestalt, an incorrect one does not. Whether or not insight is a marker of a correct answer is an important proposal to evaluate especially in light of proposals for a Eureka heuristic that may be used (and abused) to signal the quality of a new idea (Laukkonen et al., 2022).

Nonetheless, it is unclear that contrasting insight and analytical approaches to problem solution can fully address this question. Such an approach focuses on the "moment" of solution and ignores the process leading up to it. No participant starts by solving a problem through insight - it is not a strategy that can be consciously adopted. Each participant necessarily starts by applying an analytical approach; that is, to the extent that the participant focuses on the task at hand, they read the problem and labour to make sense of it. Sometimes this approach leads to success, sometimes to impasse before a traditional moment of insight and sometimes to failure. The point is that this disproportionately benefits insight solutions. Those who fail will be unlikely to label their process as insightful and if they do, it is likely that they misunderstand what it means to have an insight.

This is supported by the argument from Salvi, Bricolo, Kounios, Bowden and Beeman (2016) that insight was more likely to result in errors of omission whereas solving problems through analysis was more likely to result in errors of commission. That is that problem-solvers who report solving problems through insight were more likely to not answer at all (error of omission) rather than whose who relied on more effortful strategies were more likely to make a mistake (error of commission) presumably by guessing a tentative answer. This relies on an *overall tendency* to report insight and tells us little about the individual problem being solved – a single problem solved by insight could not also result in an error of omission: It is impossible. This suggests solving through analytical processes is more likely to encompass tentative guesses along a path of trial and error. When those trials where no answer was offered (errors of omission) are excluded while those which include tentative guesses (errors of commission) are retained (as in Salvi et al. (2016)) it leads to an exclusion asymmetry which will be more likely to endorse the proposal that insightful answers are correct.

Webb, Cropper and Little (2019) caution that more care needs to be taken over making strong claims for an accuracy effect arguing both that greater attention needs to be paid to individual dispositions towards experiencing insight and also the need to distinguish between certainty and accuracy. Of these two recommendations, the second one can be difficult to measure with many of the insight tasks used to date. Take for example the Triangle of Coins task. In this problem, 10 coins are arranged in a triangular shape pointing down: Participants must identify which three coins can be moved to rotate the orientation of the triangle such that it points up rather than down. Evidence from video based analysis (Vallée-Tourangeau, Ross, Ruffatto Rech & Vallée-Tourangeau, 2020) suggests that participants try to move various coins up. The problem requires the participant to realise that to make the triangle point *up*, the vertices must be moved *down*. Once this restructuring has occurred the problem is solved relatively simply. What is important is that moving coins up is clearly an ineffective strategy – it is impossible to think that it is correct when it is incorrect. In other words, the normative criterion demarcates sharply a correct from an incorrect answer.

This means that the correct answer and certainty in the answer are linked to a greater or lesser extent by virtue of the task employed. In addition, as argued by Webb, Laukkonen, Cropper and Little (2019), many traditional insight tasks were designed to elicit insight when solved correctly so we should not be surprised when they do exactly that. The participants can clearly know instantly whether their answer is correct even in the absence of explicit feedback and so confidence in the answer (subjective correctness) and objective correctness are difficult to disentangle. The data presented in Danek and Salvi (2020) which is a combination of two earlier papers (Danek, Fraps, von Müller, Grothe & Öllinger, 2013; Salvi et al., 2016) indicate that for problems where the answer is more ambiguous and multiple plausible responses may be possible (such as magic tricks and line drawings), over 20% of insight 'solutions' were incorrect indicating that the accuracy effect is somewhat tempered by the type of problem presented (Strickland et al., 2022).

Methodologically, to better adjudicate the plausibility and theoretical importance of the gestalt switch as a marker of accuracy, one needs a procedure that employs problems for which participants' certainty in their answers, their self-reported feelings of insight, and solution accuracy can be more clearly dissociated. An earlier paper from Danek (Danek & Wiley, 2017; also using magic tricks) suggests that the level of insight in incorrect answers is related to participants' certainty in those answers. This has been repeated across other studies and certainty is a strong component of the insight experience for both correct answers solved with insight and incorrect answers solved with insight (false insights; Danek & Wiley, 2017; Danek, Fraps, von Maller, Grothe & Ollinger, 2014). However, it may be that confidence and correctness overlap because of the form of the stimuli. For example, Stuyck et al. (2021) and Spiridonov et al. (2021) demonstrate a clear link between subjective certainty and insight using Compound Remotes Associates (CRA) but Threadgold et al. (2018) show that confidence is significantly related to a correct answer for CRA and so it is difficult to assess which is driving the insight feeling. When Danek and Wiley (2017) assessed insight through magic tricks for which it is harder to establish a normative solution, certainty was the strongest predictor of an "aha" rating for both correct and incorrect trials. This suggests that subjective correctness rather than objective correctness is the key indicator in the absence of obvious normative correct answers.

This phenomenon of false insights suggests that rather than comparing across different solution processes, the relationship between correctness, confidence and insight can be better adjudicated by considering those times when participants fail to reach a normative solution yet are still confident in their solution. Therefore, while the research reported in this paper adds to work in support of the correctness effect, it does so using stimuli which can elicit a strong feeling of confidence in a normatively incorrect answer allowing us to differentiate between the two phenomena with more care.

1.2. Stumpers as insight problems

Recently Bar-Hillel (2021; Bar-Hillel, Noah & Frederick, 2018) published a large compendium of "stumpers" which rely for their effectiveness on both the misleading initial construction and the obviousness of the answer once the problem has been restructured. A stumper is defined as "a riddle the solution to which is typically so elusive that it does not come to mind, at least initially - leaving the responder stumped" (Bar-Hillel, 2021, p. 1). For example, consider the following. "A big brown cow is lying down in the middle of a country road. The streetlights are not on, the moon is not out, and the skies are heavily clouded. A truck is driving towards the cow at full speed, its headlights off. Yet the driver sees the cow from afar easily, and avoids hitting it, without even having to brake hard. How is that possible?" (Bar-Hillel, Noah & Frederick, 2019, p. 112) .[1] They are different from most riddles in that the answer does not typically involve a compellingly intuitive but wrong answer. Instead, the individual cannot seem to come up with any answer at all for a while. It seems likely that these stumpers would require similar cognitive processes to those in insight problem solving and it may be that they would be a useful addition to the battery of insight tasks (Bar-Hillel et al., 2019). To date, there has been no research to investigate if these stumpers elicit similar feelings of insight to more traditional riddles.

Stumpers have a broader range of possible answers which could be suggested by participants - the appendix to Bar-Hillel et al. (2018) lists various inventive responses which were technically 'correct' with a level of logical contortion even if they were not the normative answer. In this way, stumpers offer the possibility which is not present in many insight tasks of disentangling certainty in the answer (subjective correctness) with the actual correct answer (objective correctness). In other words, stumpers offer the possibility of creating answer gestalts that feel right, but that correspond to normatively incorrect answers. These stimuli then offer an instrument that could help us more clearly measure the relationship between a feeling of insight and obtaining the correct answer. Investigating this relationship in complex problems such as this responds to the call by Danek and Salvi (2020) to assess to what extent the accuracy effect holds in different problem situations. Beyond this, it will support our understanding of how insight works in areas where there are no normatively correct answers available such as many of the anecdotal reports of discovery.

1.3. The current studies

We report the results from two studies that considered the relationship between the feeling of "aha" which marks an insightful answer, the feeling of certainty in the answer (subjective correctness) and the objective correctness of that answer. The use of stumpers which can generate a feeling of certainty in an incorrect answer allowed us to examine this relationship in more detail.

[1] It's daytime.

2. Study One

The primary aim of Study One was to produce normative data on the level of affective insight associated with each stumper (see similar work by Threadgold et al., 2018; Webb, Little & Cropper, 2018). In addition, we could provide the overall success rate and the average latency to solution to assess stumpers as suitable experimental stimuli. The success rates for stumpers in Bar-Hillel (2021) are a useful guide to the difficulty of the stumpers but were also collated from different research reports and almost identical stumpers appeared to elicit different rates of solution across the different experiments. For example," Alex is a blood relative of Bobbie, and Bobbie is a blood relative of Casey. Yet Alex and Casey are not blood relatives at all. How is this possible?" was solved by 44% of participants and the almost identical "Alex is a blood relative of Bobbie, and Casey is a blood relative of Bobbie. Yet Alex and Casey are not blood relatives at all. How is this possible?" was solved by 31% (see Bar-Hillel, 2021, p. 20)[2]. This may have been owing to unknown or unspecified procedural differences, such as time allocated to solve each stumper (which is not reported in Bar-Hillel, 2021). A more systematic approach was required to understand comparative normative success rates within a controlled setting and the use of a narrower time window would ensure that these tasks would be useful for other experiments.

Therefore, the first study had the primary aim of generating normative data for a set of 25 stumpers. In addition, these stimuli offered the opportunity to conduct exploratory analysis of two further hypotheses: (i) that "aha" indicates recognition of a correct answer and (ii) to assess the extent this is mediated by a feeling of certainty in the answer.

The study was preregistered on the OSF on the 16th of May 2021 (the preregistration can be found here: https://osf.io/s3jhw) The data were collected on the 19th of May 2021.

2.1. Method

2.1.1. Participants

One hundred and fifty-one participants were recruited from the online participant platform Prolific.co. We planned for a sample size which would allow us to generate meaningful normative data guided by existing studies: Threadgold et al. (2018) used 85 participants for each set of Rebus puzzles and Webb et al. (2018) reported participant numbers of 193 and 129 and 130 across the three studies. In light of these comparisons, a target sample size of 150 would yield meaningful comparisons across the stumpers; we did not compute a power analysis because the primary aim was to generate descriptive statistics.

Each participant received £7.50 in compensation. Our preregistered exclusion criteria were formulated to exclude any participant who failed both to answer 20% or more of the stumpers and to do so with log transformed response latencies faster than two standard deviations from the mean. This led to the exclusion of one participant. A further two participants were excluded for being non-native speakers. The final sample size was therefore 148 participants ($F = 66$, Other $=6$) with an average age of 26.9 ($SD = 6.0$).

2.2. Materials and measures

2.2.1. Selection of stumpers

Bar-Hillel et al. (2018) reports over 90 stumpers. Several of the stumpers were similar in structure such as "Two Italians are sharing a pizza. The older Italian is the brother of the younger Italian. But the younger Italian is not the brother of the older Italian. Explain briefly." And "Two Russians were standing in line. The taller one was the brother of the shorter one, but the shorter one was not the brother of the taller one. Explain in a few words how that is possible." In this case the research team selected one representative stumper and discarded the repetitions. This led to a set of 50 stumpers. Initial small-scale pilots suggested that some of the stumpers were either not answered at all or were answered by all participants. As those stumpers which have either floor or ceiling effects are not of interest as experimental stimuli, the final selection of stumpers was restricted to those which were solved by between 20% and 66% of pilot participants. As a result of this pilot study, and on the basis of these selection criteria, 25 stumpers were employed in Study One.

2.2.2. Additional measures

Participants also completed the CRT and CRT-v. Data for these are not directly relevant to the study here and are reported in the open access preprint Ross & Vallée-Tourangeau (2021).

2.3. Procedure

Participants were first invited to solve the stumpers in a random order. Prior to this, participants were briefed on what constituted an "aha" experience using the wording from Danek et al. (2014), itself adapted from Bowden and Jung-Beeman (2007). We used the following instruction for these judgments:

We would also like to know whether you experienced a feeling of insight when you solve each task: A feeling of insight is a kind of

[2] Alex and Casey are Bobbie's mother and father (or any two relatives from either side of Bobbie's family tree)

"Aha!" characterized by suddenness and obviousness (and often relief!)—like a revelation. In contrast, you experienced no Aha! if the solution occurs to you slowly and stepwise. As an example, imagine a lightbulb that is switched on all at once in contrast to slowly dimming it up. We ask for your subjective rating whether it felt like an Aha! experience or not, there is no right or wrong answer. Just follow your intuition.

However, we removed the sentence: "You are relatively confident that your solution is correct without having to check it" to avoid the artificially inflating the relationship between certainty and "aha" (Danek & Salvi, 2020). After reading the instructions, participants were required to indicate that they understood the feeling. Each stumper was presented for 80 s, and latency was calculated from the time the participant moved to the page displaying the stumper to when they clicked for the next page.

Participants were invited to write an answer of any description to increase the likelihood that they would offer a guess that they felt less confident in which would add to the granularity of the analysis. At the end of the problem-solving time, participants were asked to rate on a scale of 1 to 100 (a) how certain they felt in their answer and (b) whether they experienced a feeling of "aha" (Time 1 "aha"). Participants were also given the option of selecting not applicable if they had not generated an answer. Participants were then given the right answer and asked whether they were correct. All participants were asked how stuck they felt but those that self-identified as correct were also asked (a) whether they experienced a feeling of "aha" (Time 2 "aha"), (b) how surprised they felt, (c) how much they enjoyed the experience (d) how much they wanted to give up and (e) how challenged they felt.[3]

2.4. Results

All mixed effect analyses reported here were conducted using R (R Core Team, 2021) and the package lme4 was used for the mixed models (Bates, Mächler, Bolker & Walker, 2015). Following Barr, Levy, Scheepers and Tily (2013), a maximal structure was used with participants and stumper set as random intercepts and random slopes. Overall, 3700 data points were available for analysis with 148 participants and 25 stumpers. A coding error led to the "aha" at Time 2 and the questions on the impasse dimensions being omitted for one question so for those analyses, the data from 24 questions were analysed. We report the standardised coefficients and the standardised confidence intervals. The complete raw data and analysis code for both Studies One and Two are available here: https://osf.io/ujrha/?view_only=3c81e6949ee74d82a06477da0837d39b

2.4.1. Normative data

All answers were hand checked by two members of the research team to ensure that the answers the participants reported as correct mapped onto normatively correct answers. All participants offered honest assessments of their answers. Overall, 44.4% ($SD = 50.0\%$) of the stumpers were solved correctly with an average latency to a correct answer of 42.6 s ($SD = 19.3$) suggesting that they would makes suitable stimuli for repeated trials although there was a large variation between stumpers (see Appendix). The average "aha" for correct answers was 45.8 ($SD = 31.1$) and for incorrect answers was 28.0 ($SD = 28.1$). This suggests that solution of these problems can occur both with and without "aha" which makes them suitable candidates to explore different problem-solving processes (much like Rebus puzzles; Threadgold et al., 2018). The full data set is presented in the appendix.

2.4.2. Certainty

As would be expected, participants were more certain of their answer on trials when they were correct ($M = 62.8$, $SD = 30.0$) than when they were incorrect ($M = 32.2$, $SD = 29.3$). A mixed effects analysis (intercept, β −0.04 [−0.16 −0.08]) shows that having the correct answer significantly increased the certainty reported by participants, $\beta = 0.44$ [.35 - 0.53], $p < .001$. This suggests that participants showed more certainty in their answer when it was actually correct. Whether their answer was correct or not also predicted their levels of "aha" at Time 1 ($M_{correct}$ =45.8, $SD = 30.0$, $M_{incorrect} = 28.0$, $SD = 29.3$). A mixed effects analysis (intercept, β −0.02 [.09 −0.13]) demonstrates that this increase was significant, $\beta = 0.34$ [0.26 - 0.42], $p <.001$. This supports the hypothesis that insight is more likely to reflect a correct answer.

2.4.2.1. Exploratory analyses.

In order to establish which had the greater effect on the reported feeling of "aha" at Time 1, we constructed a mixed effects model with certainty as a covariate and correct/incorrect as a fixed factor (intercept, $\beta = 0.06$ [−0.03 - 0.15]). While both factors were significant contributors to the model, being certain about the answer had a greater effect on the level of "aha" reported at Time 1 $\beta = 0.59$ [.52 −0.65], $p < .001$ than whether the answer was actually correct or not, $\beta = 0.09$ [.05 - 0.13], $p < .001$

To further unpack this relationship between certainty, correctness and "aha", we divided the dataset into correct trials and incorrect trials. As we would expect from the above results, the level of certainty predicted the level of "aha" for the correct trials, (intercept, $\beta = 0.10$ [−0.02 −0.22]) $\beta = 0.51$ [.44–0.57], $p < .001$ and also for the incorrect trials, (intercept, $\beta = 0.01$ [−0.10 −0.08]) $\beta = 0.61$ [0.53–0.69], $p < .001$ demonstrating that the effect of belief in having the correct answer is important for the feeling of "aha", and this whether or not that answer is actually correct. The relationship between certainty ratings and aha ratings for each participant

[3] The additional questions were not relevant for this study and so these ratings are reported in the open access preprint Ross (2021).

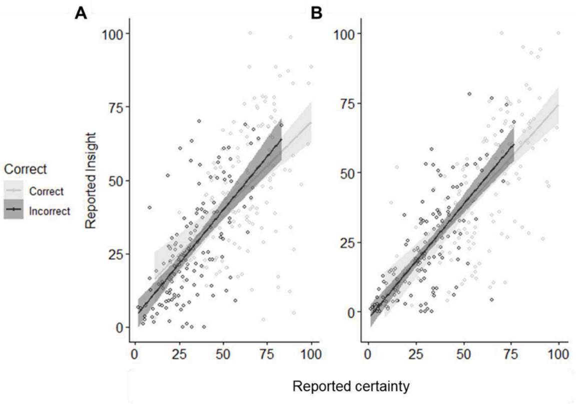

Fig. 1. The Relationship between Certainty and 'Aha' ratings for Correct and Incorrect Trials in (A) Study One and (B) Study Two. Shaded Areas Represent 95% Confidence Intervals.

is illustrated in the left panel of Fig. 1.

In light of this, it would not be unreasonable to expect "aha" at Time 2 to increase when there is increasing certainty in the answer (participants have been told they have the correct answers so certainty should be 100%) and indeed this is what we found. The overall "aha" at Time 2 ($M = 49.1$, $SD = 31.4$) was higher than at Time 1 ($M = 45.8$, $SD = 31.1$), a difference which a mixed model with "aha" as the dependant variable and time (pre/post) as a factor shows is significant, (intercept, $\beta = 0.01$ [−0.15 −0.17]), $\beta = 0.12$ [.05 −0.18], $p < .001$.

2.5. Discussion

Study One was originally designed to generate normative data for the 25 stumpers (see Appendix). The exploratory analysis suggests that rather than being linked to a correct answer, feelings of insight increased in line with certainty in the answer regardless of the objective correctness. When certainty in their answer was manipulated by telling them their answer was correct, then reported levels of insight rose. However, it could be that the reporting of "aha" at Time 2 reflected the participants' reaction to understanding the answer rather than their reported feelings of "aha". In addition, because levels of "aha" at Time 2 were not measured in those that received feedback that their answers were incorrect, we could not be sure that this difference reflected the manipulation of the feeling of certainty. Therefore, we designed and preregistered Study Two to replicate the findings of Study One and extend it in two ways. First, to ensure clarity between the feeling of "Aha" about the suggested answer and the given answer by asking two separate questions. Second, to extend the post- task questions to all participants. In addition, we were able to conduct an internal replication of the exploratory analyses from Study One.

3. Study Two

The study was preregistered on the OSF on the 19th of July 2021 (the preregistration can be found here: https://osf.io/kj3rq) The data were collected on the 6th of August 2021. It was designed as a replication and extension of Study One

3.1. Method

3.1.1. Participants

One hundred and fifty-two participants were recruited from the online participant platform Prolific.co. Each participant received £3.25 in compensation. Again, we excluded participants who failed to answer more than 20% of the questions and did so with log transformed response latencies faster than two standard deviations from the mean. This led to the exclusion of three participants. A further two participants were excluded for providing no answers to the feeling of "aha" at Time1. The final sample size was therefore 147 participants ($F = 119$, Other = 3) with an average age of 27.1 ($SD = 9.3$).

3.1.2. Materials and measures

We selected 15 stumpers based on the normative data produced in the first study to manipulate difficulty levels. Five stumpers with an overall solution rate of 9% to 22% were selected as hard stumpers, five with solutions rates of 43% to 49% were considered to be of medium difficulty and a further five with solution rates of 60% to 62% were selected as easy. We did not include the CRT or CRT-v in Study Two.

3.1.3. Procedure

The procedure was identical to Study One aside from two changes. First, in Study One only those who got the answer correct were asked (a) whether they experienced a feeling of “aha” (Time 2 “aha”), (b) how surprised they felt, (c) how much they enjoyed the experience (d) how much they wanted to give up and (e) how challenged they felt. We extended these questions to all participants. So, each participant was asked first to assess their feeling of “aha” and second how stuck they felt and then were given the answer to the question. In addition, it was possible that participants would have misread the request for an “aha” at Time 2 as relating to how they felt on learning the answer which would explain the difference in “aha” rating so we clarified this by asking two questions: ’How much did you experience a feeling of “aha” before you were told the answer?’ and ’How much did you experience a feeling of “aha” after you were told the answer?.

3.2. Results

Again, all answers were hand checked by two members of the research team to ensure that the answers the participants reported as correct mapped onto normatively correct answers. Again, all participants offered honest assessments of their answers. Overall, 47.0% (*SD* = 49.2%) of the stumpers were solved correctly with an average latency to a correct answer of 37.2 s (*SD* = 17.6) confirming that they would makes suitable stimuli for repeated trials. The average “aha” for correct answers was 44.4 (*SD* = 31.9) and for incorrect answers was 24.6 (*SD* = 26.0). The solutions rates, the mean solution latencies, and mean “aha” and certainty ratings for correct and incorrect answers are reported in Table 1.

3.2.1. Certainty

Again, participants were more certain of their answer on trials when they were correct ($M = 61.7$, $SD = 30.0$) than when they were incorrect ($M = 30.1$, $SD = 26.3$). A Mixed effects analysis (intercept, $\beta = 0.36$ [.15 - 0.57]) shows that there was a significant effect of having the correct answer on the level of certainty reported by participants, $\beta = 0.91$ [.66 −1.16], $p < .001$. This suggests that participants showed more certainty in their answer when it was actually correct. Whether their answer was correct or not also predicted their levels of “aha” at Time 1, (intercept, $\beta = 0.02$ [−0.16 −0.12]) $\beta = 0.38$ [0.27 −0.48], $p < .001$. This supports the hypothesis that insight is more likely to reflect a correct answer.

In order to establish which had the greater effect on the reported feeling of “aha” at, Time 1 we again constructed a mixed effects model with certainty as a covariate and correct/incorrect as a fixed factor (intercept, $\beta = 0.01$ [0.09 −0.10]). As before, this suggests that while both factors were significant contributors to the model, the effect of being certain about the answer has a stronger relationship with the level of “aha” reported at Time 1, $\beta = 0.64$ [.57 −0.70], $p < .001$ than whether the answer was actually correct or not, $\beta = 0.09$ [0.03 - 0.14], $p = .002$. Note that the effect here was greater than in Study One suggesting that the finding is relatively robust.

To further unpack this relationship between certainty, correctness and “aha”, we divided the dataset into correct trials and incorrect trials. As we would expect from the above results, the level of certainty predicted the level of “aha” for the correct trials, (intercept, $\beta = 0.06$ [−0.06 −0.19]) $\beta = 0.57$ [.51–0.69], $p < .001$ and for the incorrect trials, (intercept, $\beta = .{-}08$ [−0.16 −0.00]), $\beta = 0.65$ [.57–0.73], $p < .001$ suggesting that the effect of belief in having the correct answer is important for the feeling of “aha”, and this whether or not that answer is actually correct. The relationship between certainty and ‘aha’ ratings for correct and incorrect trials in Study Two is illustrated in the right panel of Fig. 2; the same strong positive relationships between certainty and ‘aha’ for *both* correct and incorrect problems observed in Study One was clearly replicated.

We predicted that “aha” at Time 2 would increase in correct trials and decrease in incorrect trials. This hypothesis was supported. The average “aha” at Time 1 was 44.5 (*SD* = 31.8) for the correct trials which increased to 50.5 (*SD* = 32.9) at Time 2; a paired samples test across correct trials suggests that this difference is significant, $t(141) = 5.26$ [10.47 – 4.75], $p < .001$. On the other hand, when participants had it confirmed that their answer was incorrect, average “aha” decreased from 24.1 (*SD* = 25.8) at Time 1 to 23.1 (*SD* = 25.9) at Time 2. This difference was not significant, $t(134) = 1.23$ [−0.75 – 3.23], $p = .220$. This replicates and extends the findings of

Table 1
Solution Rates for Stumpers in Study Two, along with the Mean Latency to Solution and Mean Certainty and Aha Ratings (with Standard Deviations) for Correct and Incorrect Solutions.

		Correct solutions			Incorrect solutions		
	Proportion correct	Latency	Certainty	Aha	Latency	Certainty	Aha
Easy	0.74	31.01 (15.74)	63.23 (29.28)	43.80 (32.78)	43.06 (21.38)	32.96 (29.02)	26.51 (27.19)
Medium	0.54	43.51 (16.78)	61.13 (29.18)	44.73 (30.41)	54.76 (18.75)	25.75 (23.83)	22.59 (24.94)
Hard	0.16	44.41 (18.26)	53.56 (30.81)	46.55 (31.92)	51.67 (20.19)	30.93 (27.62)	24.15 (25.78)

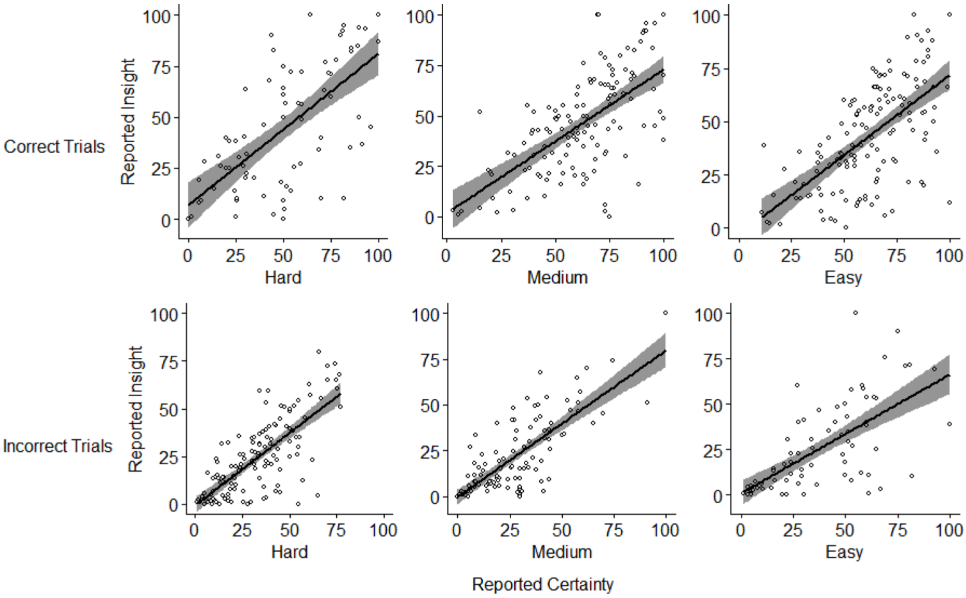

Fig. 2. The Relationship between Average Aha and Average Certainty for Correct and Incorrect Trials as a Function of Difficulty. Shaded Areas Represent 95% Confidence Intervals.

the first study which suggests that increasing certainty in the answer by assuring participants that the answer is correct increases the reported feelings of "aha". The rephrasing of the question in this second study mitigates for the possibility that the participants could be responding to the given answer rather than their suggested answer.

3.2.2. *The effect of difficulty on "aha"*

We were also interested in the effect of difficulty on the feeling of "aha" and on certainty in the correct answer. As would be expected, people solved fewer of the harder questions (16%) than either medium (52%) or easy (73%) questions. A repeated measures analysis of variance (ANOVA) shows that there was a significant effect of difficulty on the proportion of correct answers, $F(2,292) = 356.5$, $p < .001$, $\eta^2 = 0.49$ and post-hoc tests with a Bonferroni correction indicate that easy stumpers were solved more than hard stumpers ($p < .001$) and medium stumpers ($p < 0.001$) and that medium stumpers were solved more than hard stumpers ($p < .001$). Unsurprisingly perhaps, people showed more certainty in easy trials ($M = 56.1$, $SD = 20.7$) than medium trials ($M = 46.6$, $SD = 20.4$) and hard trials $M = 35.9$, $SD = 20.5$). In a repeated measures ANOVA the main effect of stumper difficulty was significant, $F(2, 276) = 72.2$, $p < .001$, $\eta^2 = 0.14$ and post hoc tests with a Bonferroni correction show that there was a significant difference ($p < .001$) across all levels of difficulty. Easy tasks ($M = 38.8$, $SD = 23.0$) elicited more insight than hard tasks ($M = 29.5$, $SD = 21.7$) and medium tasks ($M = 36.9$, $SD = 21.4$) The main effect of stumper difficulty on insight ratings was significant: F (2, 266) = 20.8, $p < .001$ $\eta^2 = 0.04$, post hoc tests with a Bonferroni correction revealed that there was a higher level of "Aha" ($p < .001$) across all levels of difficulty.

Despite these differences, the relationship between certainty in the answer and levels of "aha" represented in Fig. 2 were the same across all three levels of difficulty and whether the answers were correct or incorrect. Mixed models show that certainty was a significant predictor of "Aha" in easy tasks, (intercept, $\beta = 0.00$ [−0.18 −0.11]), $\beta = 0.58$ [.51 −0.65], $p < .001$ whereas whether the answer was correct or not was not, $\beta = 0.11$ [−0.34 −0.12], $p = .34$. This shows that for easy riddles subjective rather than objective correctness is a key predictor. This pattern was repeated with the medium difficulty tasks, where certainty was a significant predictor of "aha", (intercept, $\beta = 0.02$ [−0.08 −0.12]), $\beta = 0.70$ [.62 −0.78], $p < 0.001$ but whether the answer was correct was not, $\beta = 0.04$ [−0.16 −0.09], $p = .562$. While whether the answer was correct or not was a significant predictor of "aha" in hard trials, (intercept, $\beta = 0.26$ [.03 −0.48]) β =- 0.33 [−0.51 - −0.12], $p = .002$, certainty was still a stronger predictor, $\beta = 0.69$ [.59 −0.78], $p < .001$.

3.3. *Discussion*

Study Two replicated the findings from Study One that there is a close link between "aha" and certainty. As we found in Study One, certainty in the answer was a better predictor of the level of reported "aha" at Time 1 across both correct and incorrect answers and indeed, this relationship was stronger in this replication. Certainty and "aha" were also strongly related across the three levels of difficulty of the questions and clearly illustrated by Fig. 2. We extended Study One by asking those who had feedback that their answers were incorrect to report on their level of "aha" after feedback as well as those who had their answers confirmed. Those that got a correct answer reported significantly higher levels of "aha" at Time 2 which confirms the findings of Study One. As predicted, the levels of reported "aha" dropped when participants found out that their answer was wrong although not significantly so. We also extended the findings from Study One by conducting an analysis of difficulty. In this case when participants were faced with the easier stumpers, objective correctness ceased to be a predictor of "aha"; for these stumpers, only subjective correctness predicted 'aha'.

4. General discussion

The current studies investigated the relationship between accuracy and insight. Current research on insight uses knowledge-lean and well-structured questions to investigate the phenomenon. In these cases, it is not easily possible to have a feeling of certainty in incorrect answers and so there is necessarily a high overlap between objective and subjective correctness. Indeed, the obviousness of the answer is part of the definition of an insight task. Faith in the solution to an analytic task comes from faith in the process to reach that solution whereas in an insight task that process is opaque, and it is the answer which provides the feedback. However, claims for an objective accuracy effect beyond stimuli selected specifically to generate a feeling of insight when solved correctly must be made carefully. In short, there are many situations where there is no objective correct answer. The current study used stimuli where this relationship between objective and subjective correctness was more ambiguous, and which allowed us to investigate the relationship between "aha" and certainty. Our data support previous findings (e.g., Danek & Wiley, 2017) that suggest that certainty is robustly related to "aha" whether the answer is correct or not. These findings encourage a more nuanced interpretation of the accuracy effect. In addition, we provide normative data for 25 stumpers; these data suggest that they would useful stimuli for future problem-solving tasks.

4.1. Certainty and insight

The results here show that subjective rather than objective correctness is a bigger predictor of "aha" in problem solvers. In other words, contrary to the hypothesis put forward by Danek and Salvi (2020), insight does not reflect the quality of the solution but does reflect the certainty of the problem solver in her answer (this is clearly illustrated in Figs. 1 and 2): While participants were more certain in their answer when it was correct, the feeling of "aha" increased in line with certainty in both correct and incorrect trials in both Study One and Two and no matter what the difficulty of the problems (see Fig. 2). Indeed, with those problems which were easier to solve only certainty in the answer was a significant predictor of "aha".

By definition, the restructuring of insight problems should lead to an answer which is, with hindsight, obvious. This leads to the conflation of subjective and objective correctness in insight problem solving research. The stimuli in this study were more ambiguous in their correct answer than other stimuli typically used in insight studies where the correct answer can be tested. There was no binary answer clearly recognisable as right or wrong. Here we allowed participants to tentatively suggest answers which allowed us to begin to disentangle subjective and objective accuracy. Recognising the difference between the two is important because problems which are encountered in the more mundane flow of human experience do not have normative correct or incorrect answers. This can allow us to investigate more clearly what characteristics of an answer yield the gestalt feeling such as their aesthetic value.

Of course, it could be that the phenomenological response which accompanies a moment of "aha" increases certainty in the answer rather than that increased certainty leading to "aha". However, that "aha" increased after participants had their answer confirmed lends support to the proposal that "aha" reflects increased faith in the answer – as certainty increased so did "aha". This result is in the opposite direction to Webb et al. (2019) who found ratings of "aha" decreased for correct answers The reasons for this are unclear and point to an instability in the insight experience which requires further investigation.

Our findings also add to the growing data on the increase of the insight experience in line with an increase in the difficulty of the problem. The more difficult the problem, the more likely it is to induce an impasse and therefore trigger a more classic insight sequence (Fedor, Szathmáry & Öllinger, 2015; MacGregor, Ormerod & Chronicle, 2001; Ohlsson, 1992; Ross, 2021). We would therefore expect the results we saw, namely that the more difficult stumpers generated a greater feeling of "aha". This is in line with data from Webb et al. (2018) where a weak correlation is reported. However, other studies such as Kizilirmak et al. (2018) have found no effect of difficulty on "aha". It is plausible that the strength of "aha" in these cases is related to the strength of impasse with different stimuli eliciting different levels of impasse (Webb et al., 2018). It is a truism that "it is only easy if you know the answer" but normative and objective measures of difficulty can be problematic without knowing how difficult that solver found the problem. Without a clear measure of what makes the problem difficult (whether the impasse is merited or unmerited for one), the relationship between problem difficulty and "aha" is likely to continue to be unstable.

For a wide range of insight problems, the faith in the answer is wrought by the answer itself because it is obviously correct or not. In this study we generated that certainty by telling participants if they were correct. This suggests that it is that feeling of certainty that is important. This also reflects the process-based video analysis work on the triangle of coins that we reported in Vallée-Tourangeau et al. (2020) in which we demonstrated that all the participants who solved the problem also needed to check that the problem was solved. In that more qualitative work, we offer an example of a participant experiencing a recognisable moment of "aha" after they have constructed the answer suggesting again that "aha" comes from clear feedback that the answer is correct. It may well be that insight is a multistage process involving an insightful hunch (a 'flash of suspicion' as Chater [2018, p. 173] puts it) prior to problem solution and relief once it has been confirmed.

In summary, the data presented here suggest that the proposed intrinsic accuracy effect may need to be tested on more nuanced stimuli which allow a granularity of answer and where wrong answer are equally plausible. This is important because it is rare that a problem outside of those commonly used by problem solving researchers has a clear and unambiguous answer and false insights might be more common. Additionally, the type of solution which can elicit a clear feeling of insight might only be one which is generated by a very narrow class of tasks. This has implications both for an understanding of insight using more complex problems and for the phenomenological profile of insight itself.

4.2. Limitations

Insight problem solving is complex and unreliably evoked in laboratory studies. Any retrospective reporting of the problem-solving process necessarily elides process into the final few moments. Indeed, our data suggest and rely on the instability of post-task "aha" reporting. This observation also requires researchers to take care in assuming insight from post-task report (although see Laukkonen & Tangen, 2018), rather we suggest that other forms of measuring insight are necessary such as behavioural analyses (Ross & Vallée-Tourangeau, 2021) or physiological responses (Laukkonen et al., 2021; Salvi, Simoncini, Grafman & Beeman, 2020)

In addition, the levels of insight across the data set were lower than those reported in other norming studies such as Webb et al. (2018) or Threadgold et al. (2018). This may reflect the instructions to hazard a guess which would have encouraged participants to adopt a piecemeal, 'working towards' strategy rather than one more reliant on a gestalt shift. It may also reflect the decision to remove the explicit link between certainty and insight in the instructions.

5. Conclusion

The relationship between experimental stimuli and the phenomenon under consideration is a complex one and particularly so in the case of insight where the phenomenon is at times illusory. When it can be evoked, it is always deeply relational (Chu & MacGregor, 2011; Webb et al., 2016). Riddles and other classic insight tasks existed prior to the phenomenon of "insight" being the focus of psychological research and were co-opted because they already appeared to elicit the phenomenon: "These problems have delighted brainteaser connoisseurs for years, and most are capable of giving the solver a large dose of the 'aha!' experience" (Batchelder & Alexander, 2012, p. 49). This leads to an unproductive circularity where the correct answer to an insight problem is assumed to have been generated by insightful processes because those are the processes which are required to solve an insight problem. Insight does not lie in the problem. However, the class of problem termed insight problems do demonstrate shared characteristics such as the need for a restructuring and clear unambiguous answers (Batchelder & Alexander, 2012; Ormerod et al., 2002; Weisberg, 1995). These characteristics necessarily constrain the conclusions that can be drawn. The data here suggest that it is this latter characteristic and the certainty which accompanies a correct answer of this kind which may be important in understanding the phenomenology of insight.

CRediT authorship contribution statement

Wendy Ross: Conceptualization, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Frédéric Vallée-Tourangeau:** Writing – original draft, Writing – review & editing.

Data availability

All data is avaliable at https://osf.io/ujrha/?view_only=220452a5be314d7b9166f950cd05657e

Appendix: Normative data for stumpers from study one split in terms of correct and incorrect trials (Standard deviations are reported in brackets)

Normative data for stumpers from study one split in terms of correct and incorrect trials (Standard deviations are reported in brackets)

		Correct						Incorrect					
	Proportion Correct	Latency		Certainty		Aha		Latency		Certainty		Aha	
Hardy Pyle was bragging about his church's baseball team. He said, Three of our players hit home runs and two of those home runs were hit with the bases loaded. Our guys won 9 to 0 and not a single man crossed home plate. How was this possible?	0.09	57.60	(18.98)	53.93	(31.94)	49.92	(32.71)	52.51	(19.67)	19.28	(25.00)	24.04	(32.37)
Individual bus rides cost one dollar each. A card good for five rides costs five dollars. A first-time passenger boards the bus alone, and hands the driver five dollars, without saying a word. Yet the bus driver immediately realizes, for sure, that the passenger wants the card, rather than a single ride and change. How come?	0.12	59.43	(16.64)	60.28	(29.68)	54.35	(26.92)	64.15	(15.64)	27.26	(24.26)	28.01	(29.13)
Dan and Daphne had a fight just before Valentine's Day, and Daphne was hoping that Dan would try to make up. On Valentine's Day, Dan sent Daphne a big Valentine's Day Special chocolate box from her favourite chocolatier ("La Maison du Chocolat"). Explain briefly why the brokenhearted Daphne furiously tossed it into the garbage.	0.15	49.30	(23.29)	47.60	(33.61)	33.37	(28.42)	61.55	(16.36)	27.52	(26.15)	23.82	(25.86)
Stephanie was reading a book in her bedroom when suddenly all the lights went off. The house was now pitch dark. Explain in a few words how come Stephanie calmly went on reading as before.	0.17	34.37	(17.26)	74.16	(24.48)	54.88	(34.55)	44.74	(18.45)	47.76	(29.08)	33.74	(27.06)
Jorge, whose vision is 20/20, stood, with eyes wide open, looking directly at a large modern painting, hanging at eye alevel on a wall just 2 yards away, with nothing occluding it. Jorge looked and looked but could not see the painting. Explain briefly how that is possible.	0.22	51.35	(20.09)	50.06	(28.52)	44.58	(28.85)	54.04	(17.71)	32.17	(26.50)	28.15	(24.93)
Two Russians were standing in line. The taller one was the brother of the shorter one, but the shorter one was not the brother of the taller one. Explain in a few words how that is possible.	0.24	40.97	(17.00)	78.83	(25.53)	61.83	(30.84)	59.94	(19.10)	20.89	(24.30)	20.39	(27.62)
There is a certain prayer during which it is customary to kneel. At a recent church gathering, all present were kneeling. Explain briefly why Maria was not.	0.28	40.16	(17.70)	62.61	(34.45)	47.07	(32.38)	39.08	(17.29)	39.90	(31.07)	26.56	(26.69)
Who has both feet on the beach while flying?	0.32	38.96	(21.62)	41.72	(27.90)	42.78	(28.79)	50.78	(22.30)	20.98	(26.21)	23.93	(29.84)
A hungry horse is tied by its neck to a 10-metre-long chain. A bale of hay is 13.8 m away from it. Explain briefly how the horse reaches the hay with the chain intact.	0.43	48.51	(17.21)	57.06	(29.16)	43.08	(31.19)	60.06	(18.63)	29.33	(25.12)	23.71	(22.03)
Dame Dora owns an Old Masters painting in a heavy gilded frame. The cord for hanging the painting, as old as the painting itself, is made of thick 3-ply hemp, and is somewhat frayed. Dame Dora was thinking of replacing it. But before she could, a couple of hungry little mice invaded her mansion. Sneaking behind the painting, they chewed right through the cord. For a while nobody noticed because the painting didn't budge. Explain the painting's stability briefly.	0.43	56.59	(16.84)	44.92	(31.56)	33.05	(25.61)	64.31	(16.91)	25.02	(26.15)	21.54	(25.31)
A man in town married 20 women in the town. He and the women are still alive, and he has had no divorces. He is not a bigamist and is not a Mormon and yet he broke no law. How is that possible?	0.44	40.65	(20.33)	75.83	(26.12)	64.56	(27.33)	58.65	(18.14)	25.78	(30.31)	21.67	(25.72)
Denise is a pretty good tennis player. She made a bet that she could hit a regular tennis ball, send it flying off in the air, and after a bit, it would turn around 180° and fly right back to her – without making contact with any other object on its way. She won the bet. Explain how in a few sensible words.	0.46	54.92	(18.18)	69.26	(30.83)	53.21	(31.53)	60.74	(17.89)	29.97	(27.08)	26.62	(27.16)
Tom broke his arm badly, and it was in a cast for weeks. When the cast was removed, he trained as follows: He extended his arm to the side, straight,	0.49	53.29	(17.12)	56.39	(30.01)	44.06	(29.61)	67.20	(14.77)	24.91	(28.73)	27.17	(31.89)

(*continued on next page*)

(*continued*)

and while holding a small potato bag, maintained this position for as long as he could. Once he could keep it that way for a whole minute, the small bag was replaced by a medium bag, and the exercise repeated. Once he could hold the medium bag for a full minute, it was replaced by a large bag. As soon as Tom could hold a large potato bag that way for an entire minute, one potato was added to the bag. Tom's arm collapsed almost immediately. How come?													
It is raining cats and dogs (i.e., it is pouring). Four people try to squeeze underneath one small umbrella. Explain briefly how nobody gets wet.	0.50	42.04	(19.98)	45.96	(28.85)	31.66	(25.39)	49.77	(19.68)	34.24	(30.19)	31.92	(31.21)
While walking on the newly paved black asphalt road to her home, Jen accidentally dropped her small black leather clutch. The lights on the new road had not been turned on yet, the moon was not out, and Jen did not have on her a flashlight, matches, mobile phone, or any other means of lighting. Explain briefly how Jen nonetheless saw her clutch immediately.	0.51	49.00	(17.79)	67.45	(27.14)	48.43	(28.44)	62.41	(16.39)	27.18	(24.80)	24.51	(25.35)
Hardy Pyle was washing windows on a high-rise office building when he fell off his 60-foot ladder onto the concrete sidewalk below. Incredibly, he did not injure himself in any way. How was this possible?	0.52	44.87	(19.17)	59.23	(27.33)	41.03	(26.55)	55.51	(17.17)	37.73	(30.21)	28.65	(27.66)
Maxine walked for 200 m directly on the surface of a lake, without sinking, without any devices, and without getting any clothing wet. Explain in a few words how she managed this.	0.53	29.82	(14.49)	80.96	(22.48)	49.96	(33.97)	47.10	(20.16)	47.92	(33.09)	36.28	(31.87)
What gets wet as it dries?	0.59	29.28	(17.13)	73.47	(28.04)	49.63	(34.81)	39.73	(23.74)	36.93	(32.09)	34.05	(30.92)
On Christmas Day, at the stroke of midnight, David walked out to his own back yard, on a dare. He was stark naked: no shoes, no socks, no sweater, no coat, no hat, no scarf - nothing. He stood out there with arms outstretched, singing Christmas carols, for 5 whole minutes. When he came back indoors, he wasn't the least bit cold. Explain briefly.	0.60	50.78	(15.42)	60.51	(29.63)	41.22	(31.10)	61.50	(17.27)	27.68	(28.50)	26.32	(27.02)
Bob's driver's license was recently revoked, following a string of severe traffic violations. Just a few days later, a cop spotted the unlicensed Bob yet again, entering a one-way street against the direction of the traffic. This was the same cop who had cited Bob before. Explain briefly how come the cop did not stop him, and just gave him a smile.	0.60	39.14	(16.32)	71.45	(25.79)	49.81	(29.90)	57.85	(16.90)	35.21	(31.27)	38.52	(31.19)
A clerk at a butcher shop stands five feet ten inches tall and wears size 13 sneakers. What does he weigh?	0.62	32.46	(16.75)	67.67	(29.75)	62.20	(30.97)	51.63	(21.01)	20.16	(22.97)	20.04	(26.48)
Farmer Joe eats two fresh eggs from his own farm for breakfast every day. Yet there are no chickens on his farm. Where does Farmer Joe get his eggs?	0.62	43.83	(19.90)	44.36	(30.89)	29.28	(26.71)	42.14	(20.99)	47.17	(33.06)	35.38	(27.1)
Laura took a multiple-choice test. She barely speaks, reads, or understands English, and had nobody who could translate for her. Explain briefly how Laura scored nearly 100% on the test, completely legitimately.	0.62	38.64	(17.80)	65.25	(27.67)	45.04	(28.36)	44.30	(19.01)	37.96	(28.45)	24.08	(26.31)
A very tall man was holding up a wine decanter way above his head. He let go of it, and it dropped to the carpet he was standing on. Explain briefly how not a single drop of wine was spilled.	0.73	33.90	(17.20)	71.33	(25.32)	46.66	(32.13)	57.06	(20.15)	41.64	(26.89)	33.03	(31.36)
Barney Dribble is carrying a pillowcase full of feathers. Hardy Pyle is carrying three pillowcases the same size as Barney's, yet Hardy's load is lighter. How can this be?	0.83	42.67	(14.94)	63.10	(27.19)	40.94	(30.47)	48.72	(17.46)	41.77	(31.04)	32.47	(28.27)

References

Bar-Hillel, M. (2021). Stumpers: An annotated compendium*. *Thinking & Reasoning, 27*(4), 536–566. https://doi.org/10.1080/13546783.2020.1870247

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

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Batchelder, W. H., & Alexander, G. E. (2012). Insight problem solving: A critical examination of the possibility of formal theory. *The Journal of Problem Solving, 5*(1). https://doi.org/10.7771/1932-6246.1143

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 10/gcrnkw.

Bowden, E. M., & Jung-Beeman, M. (2003). Aha! Insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin & Review, 10*(3), 730–737. https://doi.org/10.3758/BF03196539

Chronicle, E. P., MacGregor, J. N., & Ormerod, T. C. (2004). What makes an insight problem? The roles of heuristics, goal conception, and solution recoding in knowledge-lean problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 14–27. https://doi.org/10.1037/0278-7393.30.1.14

Chu, Y., & MacGregor, J. N. (2011). Human performance on insight problem solving: A review. *The Journal of Problem Solving, 3*(2). https://doi.org/10.7771/1932-6246.1094

Danek, A. H., Fraps, T., von Maller, A., Grothe, B., & Ollinger, M. (2014). It's a kind of magic: What self-reports can reveal about the phenomenology of insight problem solving. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.01408

Danek, A. H., Fraps, T., von Müller, A., Grothe, B., & Öllinger, M. (2013). Aha! experiences leave a mark: Facilitated recall of insight solutions. *Psychological Research, 77*(5), 659–669. https://doi.org/10.1007/s00426-012-0454-8

Danek, A. H., & Salvi, C. (2020). Moment of truth: Why Aha! experiences are correct. *The Journal of Creative Behavior, 54*(2), 484–486. https://doi.org/10.1002/jocb.380

Danek, A. H., & Wiley, J. (2017). What about false insights? Deconstructing the Aha! experience along Its multiple dimensions for correct and incorrect solutions separately. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.02077

Fedor, A., Szathmáry, E., & Öllinger, M. (2015). Problem solving stages in the five square problem. *Frontiers in Psychology, 6*, 1050. https://doi.org/10.3389/fpsyg.2015.01050

Friedlander, K. J., & Fine, P. A. (2018). The penny drops": Investigating insight through the medium of cryptic crosswords. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00904

Hill, G., & Kemp, S. M. (2018). Uh-Oh! What have we missed? A qualitative investigation into everyday insight experience. *The Journal of Creative Behavior, 52*, 201–211. https://doi.org/10.1002/jocb.142

Kizilirmak, J. M., Gallisch, N., Schott, B. H., & Folta-Schoofs, K. (2021). Insight is not always the same: Differences between true, false, and induced insights in the matchstick arithmetic task. *Journal of Cognitive Psychology, 33*(6–7), 700–717. https://doi.org/10.1080/20445911.2021.1912049

Kizilirmak, J. M., Serger, V., Kehl, J., Öllinger, M., Folta-Schoofs, K., & Richardson-Klavehn, A. (2018). Feelings-of-Warmth increase more abruptly for verbal riddles solved with in contrast to without Aha! Experience. *Frontiers in Psychology, 9*, 1404. https://doi.org/10.3389/fpsyg.2018.01404

Laukkonen, R., Ingledew, D. J., Grimmer, H. J., Schooler, J. W., & Tangen, J. M. (2021). Getting a grip on insight: Real-time and embodied Aha experiences predict correct solutions. *Cognition and Emotion, 35*(5), 918–935. https://doi.org/10.1080/02699931.2021.1908230

Laukkonen, R., & Tangen, J. M. (2018). How to detect insight moments in problem solving experiments. *Frontiers in Psychology, 9*, 282. https://doi.org/10.3389/fpsyg.2018.00282

Laukkonen, R., Webb, M. E., Salvi, C., Tangen, J. M., Slagter, H. A., & Schooler, J. (2022). *Eureka Heuristic: How feelings of insight signal the precision of a new idea* [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/ez3tn

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 176–201. https://doi.org/10.1037//0278-7393.27.1.176

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane, & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking* (pp. 1–44). Harvester Wheatsheaf.

Ormerod, T. C., MacGregor, J. N., & Chronicle, E. P. (2002). Dynamics and constraints in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 791–799. https://doi.org/10.1037//0278-7393.28.4.791

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Ross, W. (2021). *Feeling stumped: Investigating dimensions of impasse* [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/4zhse. https://web.archive.org/web/20211209194922/https://psyarxiv.com/4zhse/

Ross, W., & Vallée-Tourangeau, F. (2021). Kinenoetic analysis: Unveiling the material traces of insight. *Methods in Psychology, 5*. https://doi.org/10.36850/e4

Salvi, C., Bricolo, E., Kounios, J., Bowden, E. M., & Beeman, M. (2016). Insight solutions are correct more often than analytic solutions. *Thinking & Reasoning, 22*(4), 443–460. https://doi.org/10.1080/13546783.2016.1141798

Salvi, C., Simoncini, C., Grafman, J., & Beeman, M. (2020). Oculometric signature of switch into awareness? Pupil size predicts sudden insight whereas microsaccades predict problem-solving via analysis. *NeuroImage, 217*, Article 116933, 10/gmndq6.

Spiridonov, V., Loginov, N., & Ardislamov, V. (2021). Dissociation between the subjective experience of insight and performance in the CRA paradigm. *Journal of Cognitive Psychology, 33*(6–7), 685–699. https://doi.org/10.1080/20445911.2021.1900198

Strickland, T., Wiley, J., & Ohlsson, S. (2022). Hints and the Aha-Accuracy effect in insight problem solving. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society* (pp. 3209–3215).

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*, Article 103055. https://doi.org/10.1016/j.concog.2020.103055

Threadgold, E., Marsh, J. E., & Ball, L. J. (2018). Normative data for 84 UK English rebus puzzles. *Frontiers in Psychology, 9*, 2513. https://doi.org/10.3389/fpsyg.2018.02513

Vallée-Tourangeau, F., Ross, W., Ruffatto Rech, R., & Vallée-Tourangeau, G. (2020). Insight as discovery. *Journal of Cognitive Psychology, 33*(6–7), 718–737. https://doi.org/10.1080/20445911.2020.1822367

Webb, M. E., Cropper, S. J., & Little, D. R. (2019). Aha!" is stronger when preceded by a "huh?": Presentation of a solution affects ratings of aha experience conditional on accuracy. *Thinking & Reasoning, 25*(3), 324–364. https://doi.org/10.1080/13546783.2018.1523807

Webb, M. E., Laukkonen, R. E., Cropper, S. J., & Little, D. R. (2019). Commentary: Moment of (Perceived) truth: Exploring accuracy of Aha! experiences. *The Journal of Creative Behavior, jocb.433*. https://doi.org/10.1002/jocb.433

Webb, M. E., Little, D. R., & Cropper, S. J. (2016). Insight is not in the problem: Investigating insight in problem solving across task types. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01424

Webb, M. E., Little, D. R., & Cropper, Simon. J. (2018). Once more with feeling: Normative data for the aha experience in insight and noninsight problems. *Behavior Research Methods, 50*(5), 2035–2056. https://doi.org/10.3758/s13428-017-0972-9

Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In R. J. Sternberg, & J. Davidson (Eds.), *The nature of insight* (pp. 157–196). MIT Press.