COGNITION AND EMOTION
2023, VOL. 37, NO. 2, 329–338
https://doi.org/10.1080/02699931.2023.2187352

Routledge
Taylor & Francis Group

The illusion of insight: detailed warnings reduce but do not prevent false "Aha!" moments

Hilary J. Grimmer[a], Jason M. Tangen[a], Anna Freydenzon[c] and Ruben E. Laukkonen[b]

[a]School of Psychology, The University of Queensland, Saint Lucia, Australia; [b]School of Psychology, Southern Cross University, Lismore, Australia; [c]Institute of Molecular Bioscience, The University of Queensland, Saint Lucia, Australia

ABSTRACT

False "Aha!" moments can be elicited experimentally using the False Insight Anagram Task (FIAT), which combines semantic priming and visual similarity manipulations to lead participants into having "Aha!" moments for incorrect anagram solutions. In a preregistered experiment ($N = 255$), we tested whether warning participants and explaining to them exactly how they were being deceived, would reduce their susceptibility to false insights. We found that simple warnings did not reduce the incidence of false insights. On the other hand, participants who were given a detailed explanation of the methods used to deceive them experienced a small reduction in false insights compared to participants given no warning at all. Our findings suggest that the FIAT elicits a robust false insight effect that is hard to overcome, demonstrating the persuasive nature of false insights when the conditions are ripe for them.

ARTICLE HISTORY
Received 20 June 2022
Revised 24 February 2023
Accepted 1 March 2023

KEYWORDS
Insight; Aha; warnings; phenomenology

An insight moment is often taken to be a signal of accuracy as solutions reached via an "Aha!" experience feel immediately true and are likely to be correct (Bowden & Jung-Beeman, 2003; Kounios & Beeman, 2009; Laukkonen et al., 2021; Salvi et al., 2016). However, "Aha!" experiences do not always predict correct solutions (Webb et al., 2019), and recent work has shown that false "Aha!" moments can be elicited artificially in laboratory experiments. The False Insight Anagram Task (FIAT) reliably leads participants to have insight experiences for incorrect anagram solutions by priming a semantic category and presenting anagrams that look like a primed associate (Grimmer et al., 2022a). In the current study, we investigated whether the FIAT effect persists despite warning participants of potential deception or explaining the effect in detail.

The FIAT was inspired by the classic Deese, Roediger, and McDermott (DRM) paradigm (Roediger & McDermott, 1995), which elicits false memories of words that are semantically related to a list of studied primes. In the FIAT, participants are presented with lists of 10 semantically related words (e.g. REMEMBER, SIGNIFICANT, HONOUR, TRIBUTE, MEMORIAL, STATUE, etc.) and told to remember them for a later test. After studying the words, they are asked to solve a series of anagrams, one of which looks like a semantic associate from the studied list when scrambled (e.g. MEMUNOMT). Participants tend to solve this *primed lure* anagram with the incorrect semantic associate (i.e. MONUMENT) and often report experiencing an "Aha!" moment when they arrive at this incorrect solution (the correct solution is MOMENTUM). This procedure reliably elicits "Aha!" moments for incorrect solutions at a far greater rate than the level of false insights experienced for control anagrams (Grimmer et al., 2022a), with the effect driven by a combination of semantic priming

CONTACT Hilary J. Grimmer hilary.grimmer@uqconnect.edu.au School of Psychology, The University of Queensland, Saint Lucia, 4072, Australia

and visual similarity (anagrams must be scrambled in a configuration that looks similar enough to the intended false solution).

Having established the utility of the FIAT for eliciting false insights, we tested whether a variety of individual differences might predict people's tendency to experience false insights (Grimmer et al., 2022b). 200 participants completed the FIAT along with several measures of psychosis proneness and thinking style, given that these measures have been shown to correlate with the DRM effect (Watson et al., 2005). Mirroring the findings of Nichols and Loftus (2019), measures of psychosis proneness and thinking style did not predict false insights on the FIAT. These results suggest that the process underlying the FIAT may reflect a more general tendency of human cognition that is not unique to any particular type of person.

Based on the above findings, here we aimed to test whether the FIAT operates through an automatic process that is largely outside of conscious control or whether it can be disrupted with advanced warning. Many cognitive biases are surprisingly robust to warning participants about the effect, incentivising them to avoid it, or teaching them exactly how it occurs. For example, the illusory truth effect, which occurs when people rate previously seen statements as more true than unseen statements, is attenuated but does not disappear when people are forewarned (Jalbert et al., 2020; Nadarevic & Aßfalg, 2017). Similarly, the revelation effect—a bias towards remembering, preferring, and believing statements containing correctly-solved anagrams—also persists when participants are warned about the effect and given feedback on the accuracy of their memory judgements (Aßfalg & Nadarevic, 2015). The DRM paradigm has also been tested in numerous debiasing contexts. Many of these reveal that false memories emerge even when people try to avoid them but can be reduced by sufficiently detailed warnings (Anastasi et al., 2000; Gallo et al., 1997; McDermott & Roediger, 1998; Starns et al., 2007).

The mixed results of these earlier studies are likely due in some part to the variety of ways they have operationalised and employed warnings in attempt to reduce distortions of memory or judgement. For example, Jalbert and colleagues (2020) found that warnings only reduced the illusory truth effect when presented prior to exposure, not prior to the test, suggesting that warnings change how stimuli are processed (i.e. the encoding phase), not by encouraging doubt when interpreting feelings of familiarity and truth (i.e. the metacognitive appraisal phase). McCabe and Smith (2002) also found that warnings presented before the DRM task reduced false memories. Thus, warnings seem to reduce false memories due to changes in the encoding phase.

Other studies show that the more detailed the warnings are, the greater the extent to which they reduce the incidence of false memories (Beato & Cadavid, 2015; Gallo et al., 1997). Some previous studies have used general warnings, such as Jalbert and colleagues (2020), who warned participants that they would encounter false information, rather than explaining and warning against the illusory truth effect in detail. Others have provided an explicit warning against a particular psychological effect, by informing participants about the particular errors they are likely to make (e.g. Calvillo & Parong, 2016; McDermott & Roediger, 1998). Gallo and colleagues (1997) found that the DRM effect was robust to three warning conditions (detailed, simple, or no warning). However, there was a reduction in errors when comparing participants given detailed warnings to no warnings. Indeed, if warnings operate at the level of encoding, it makes sense that more specific instructions to avoid misleading cues would lead to a greater reduction in errors, as explicit warnings would tell participants exactly what they should be wary of during the study phase. For example, if a participant completing the DRM task has been told how the DRM effect works, they could begin the experiment intending to only remember words based on their appearance, not their semantic category, as they know this is an unreliable strategy.

In predicting how the FIAT will respond to warnings, we must also consider the different cognitive mechanisms underlying the effects used in previous warning studies, and how these mechanisms may operate in the FIAT. The illusory truth effect, for example, works by increasing the processing fluency of a stimulus (Newman et al., 2020) – the more we read a statement, the easier we will process it the next time we encounter it. This increased ease of processing makes the statements feel more familiar and trustworthy than unseen statements (Unkelbach & Greifeneder, 2018). The revelation effect is also thought to be a product of processing fluency (Aßfalg & Bernstein, 2012; Bernstein et al., 2002). Solving an anagram correctly increases the perceptual fluency of perceiving that solution, making it more readily recalled, preferred, and perceived as true (Bernstein et al., 2002; Kronlund & Bernstein, 2006). On the other hand, the DRM effect is driven by the shared semantic category of the studied words extending into the reconstructive process of memory via associative processing (Otgaar et al., 2019; Zhu et al., 2013).

The FIAT requires a combination of visual similarity and semantic priming to produce false insights (Grimmer et al., 2022a). In Grimmer et al. (2022a), the effect of both visual similarity and semantic priming disappeared when either manipulation was removed. This finding demonstrates that the FIAT is partially driven by semantic priming like the DRM effect. However, unlike the DRM, processing fluency may also be involved, as the visual similarity manipulation may make the anagrams easier to process and solve.

Based on the findings described above, we reasoned that false insights would persist despite warnings, but anticipated that more explicit warnings would reduce them more than general cautioning against errors.[1] This is because warnings about a particular effect seem to change how information is processed at the encoding phase of the task in both the DRM and illusory truth effects (Jalbert et al., 2020; McCabe & Smith, 2002). We would therefore expect that detailed warnings would give participants better knowledge about what errors they ought to avoid and improve their efforts to avoid them (Gallo et al., 1997). Taking a similar approach to Gallo et al. (1997), we warned one group of participants to be cautious of error when completing the FIAT and gave another group of participants the same warning along with a detailed explanation of the false insight effect and a demonstration of how the FIAT produces false insights. We predicted that warning participants against giving incorrect answers would lead to a decreased rate of false insights. However, given the findings of earlier research (Gallo et al., 1997) we also predicted that making participants fully aware of how the false insight is being triggered, and instructing them to avoid it, would reduce false insights more than a general warning.

Method

Open practice statement

This experiment is preregistered on the Open Science Framework. The experimental design, materials, video instructions, analysis scripts, and data are available at: https://osf.io/y2kxr/.

Participants

We simulated and analysed the results from 2000 datasets based on the mean differences reported in Gallo et al. (1997) between the three conditions. This entire analysis is available in our materials folders (https://osf.io/y2kxr/files/). These simulations revealed that a three-way, between-groups design with 63 participants in each condition would be sufficient to detect our predicted differences between the instruction conditions in all 2000 of these simulations (100%). By decreasing the mean differences on the FIAT effect between the three instruction types to derive the smallest effect size of interest, which was $\eta^2_G = .01$ (Lakens et al., 2018) our simulation revealed that we could still detect a significant interaction between Instruction Condition and Anagram Type in 1600 out of 2000 simulated datasets (80%). Based on these results, we used a sample of 255 native-English speaking participants (127 males, 127 females, 1 non-specified, mean age = 35.95 years) from Prolific Academic who were paid $5 for their participation.

Design & materials

The experimental materials were identical to those in Grimmer et al. (2022b). Participants were presented with a list of ten semantically related words associated with a certain category (e.g. PATRIOTIC, POPULIST, IDEOLOGY, PRIDE, RACIST, HOMELAND, CULTURE, NAZI, FREEDOM, RIGHT-WING). After reading the list, participants were presented with two anagrams to solve in random order. One anagram was a word either from the list (e.g. TAPRIOITC – PATRIOTIC) or from the same semantic category (e.g. AUTRORIATAIHN – AUTHORITARIAN). The other anagram was the *primed lure*, which is a word that looks like another associate, but did not appear on the list, or share a semantic association with it. For example, NLATONALIITS looks like NATIONALISTS when scrambled, but the correct solution is actually INSTALLATION. These primed lures induce more false insights than the non-lure anagrams by virtue of their visual similarity to the primed associate (Grimmer et al., 2022a).

We used a 2 (Anagram Type: Primed Lure, Other) x 3 (Instruction Type: Control, Warning, Warning + Explanation) mixed design with the between-groups variable being Instruction Type. The focal independent variable was the type of instructions (e.g. warnings) given to participants. In addition to the original instructions used in Grimmer et al. (2022a), we created two new conditions with different instructional videos (see the OSF for the videos and Appendix A for the full transcripts) designed to reduce or eliminate the false insight effect of the primed lure anagrams. By creating these three conditions, we can observe whether there is an effect of awareness in preventing false insights. The different conditions are described in Table 1 below.

Procedure

The experiment was conducted online through Qualtrics, as in Grimmer et al. (2022a). Participants were randomly assigned to one of the three instruction conditions presented in Table 1. After the instructions were presented, participants were shown a practice trial of the task. They were then asked to answer a multiple-choice comprehension check question about the study based on the instructions to confirm they had understood them. depending on which condition participants were assigned to (see Appendix B). If participants failed this comprehension check, they were asked to attempt the question again until they had selected the right response. Once they passed the check, participants were ready to begin the experiment.

Table 1. Three instruction conditions.

Control	Warning	Warning + Explanation
The first group served as a control group given no warning about the possibility of experiencing false insights. Participants watched a video with the standard FIAT instructions in which they are told to study the list and remember as many as possible, *Aim*: Replicate the findings from Grimmer et al. (2022a).	Participants warned that the experiment contained a trick that could lead them to solve some anagrams incorrectly without revealing the nature of the effect completely. They watched a video with the same instructions as the control group and a brief warning that half of their solutions may be incorrect. *Aim*: Test whether cautioning participants against providing incorrect solutions would make them more careful when solving the anagrams.	Participants were given the same warning as the previous condition along with a full explanation of the FIAT paradigm. The video also contained a visual demonstration of how the FIAT elicits false insights. *Aim*: Our best attempt to eliminate the false insight effect by not just raising participants' caution but telling them exactly what to be cautious of.

The experimental task followed the protocol from Grimmer et al. (2022a). Participants were instructed to study the list of words and try to recall them during a subsequent memory task. After studying each list of words, participants solved two anagrams, one of which was a primed lure anagram (Grimmer et al., 2022a) which was designed to resemble a word that is semantically related to the list of studied words but did not actually appear on the list. The other anagram served as a control to ensure participants could not detect or predict which anagrams were tricks, so it was either a scrambled word taken from the list (a presented target), or a scrambled word that was semantically related to the items on the list (a primed target).

When participants solved the anagram, they clicked an arrow and typed their solution into a field that appeared on the screen. After providing their solution, they indicated whether they had an insight moment. If participants responded "yes", they were asked to rate the intensity of their "Aha!" moment on a scale of 1 ("weak") to 10 ("strong"). After solving the two anagrams, which were presented in random order, participants were then asked to list as many words from the original study list as they could remember.[2] This process was repeated another 11 times. After completing the experiment, participants were debriefed and reimbursed for their time.

On trials where participants responded "yes" to experiencing an insight moment, and the solution they provided was incorrect, this was coded as a false insight. As in our previous studies (Grimmer et al., 2022a), solutions were only regarded as incorrect when either the intended (primed) solution (e.g. MONUMENT or NATIONALISTS) or a similar looking but unrelated word was provided (e.g. MONUMENTAL or NATIONALITIES).

Results

The proportion of trials with false insights reported for the Control, Warning, and Warning + Explanation conditions are depicted in Figure 1.

To test whether our warnings effectively reduced false insights, we conducted a 2 (Anagram Type: Primed Lure, Other) x 3 (Instruction Condition: Control, Warning, Warning + Explanation) mixed ANOVA with Instruction Condition as the between-participants variable. As expected, this analysis

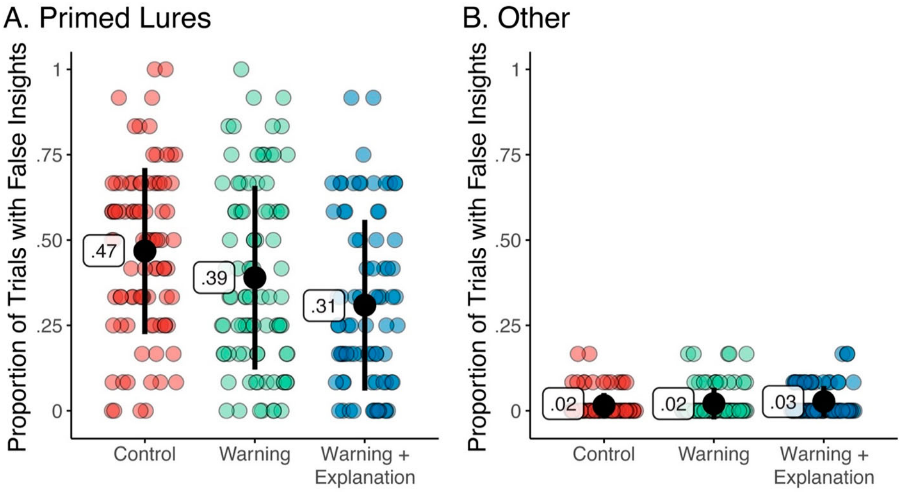

Figure 1. Proportion of Trials With False Insights Reported in Each Instruction Condition for each Anagram Type. Note. Data points illustrate the proportion of trials with false insights recorded for each participant. Means and standard deviations are shown in black.

revealed a significant main effect of Anagram Type, $F(1,252) = 568.33$, $p < .001$, $\eta^2_G = .507$, a significant main effect of Instruction Condition, $F(2,252) = 6.21$, $p = .002$, $\eta^2_G = .026$, and a significant interaction between Anagram Type and Instruction Condition, $F(2,252) = 10.03$, $p < .001$, $\eta^2_G = .035$.

We followed up the main effect of Anagram Type to confirm that the primed lure anagrams were having their intended effect. As predicted, a Tukey *t*-test revealed that the primed lure anagrams elicited significantly more false insights ($M = 0.39$, $SD = 0.26$) than the control anagrams ($M = 0.02$, $SD = 0.04$), $t(252) = 23.84$, $p < .001$, CI = 0.35, 0.40.

Next, we ran a series of Tukey *t*-tests to follow-up the focal significant effect of Instruction Condition. Contrary to our predictions, there was no difference in rates of false insights between participants in the Control condition ($M = 0.24$, $SD = 0.29$) and participants who were given a warning ($M = 0.21$, $SD = 0.27$), $t(252) = 1.82$, $p = .182$, CI = −0.01, 0.08. Only participants who were given a warning and a full explanation of the effect ($M = 0.17$, $SD = 0.23$) showed significantly fewer false insights than participants who were given no warning at all, $t(252) = 3.53$, $p = .002$, CI = 0.02, 0.12. No difference emerged between participants in the Warning condition and the Warning + Explanation condition, $t(252) = -1.79$, $p = .176$, CI = −0.08, 0.01. Taken together, these findings suggest that simple warnings are not sufficient to prevent the false insight effect from occurring.

We then investigated the interaction between Anagram Type and Instruction Condition by comparing the rates of false insights for each instruction condition for each Anagram Type. We began with the primed lures, as these were the anagrams that elicited the majority of false insights. Figure 1, Panel A illustrates that, as expected, a difference between the three types of instructions as revealed by a significant main effect of Instruction Condition in a one-way ANOVA, $F(2,252) = 8.14$, $p < .001$, $\eta^2_G = .061$. We followed up this result by conducting a series of Tukey pairwise comparisons, which again revealed that participants in the Warning condition did not experience significantly fewer false insights ($M = 0.39$, $SD = 0.27$) than those in the Control condition ($M = 0.47$, $SD = 0.24$), $t(252) = 2.03$, $p = .108$, CI = −0.01, 0.17. There was also no difference between rates of false insights in the Warning condition compared to the Warning + Explanation condition ($M = 0.31$, $SD = 0.25$), $t(252) = -2.04$, $p = .104$, CI = −0.17, 0.01. However, participants who received a full explanation of the paradigm experienced fewer false insights than those who were given no warning at all, $t(252) = 4.03$, $p < .001$, CI = 0.07, 0.25. We then ran the same analysis looking at the control anagrams (see Figure 1, Panel B), which revealed no effect of Instruction Condition, $F(2,252) = 1.61$, $p = .201$, $\eta^2_G = .013$.

Finally, we compared the rates of false insights between primed lures and control anagrams for each type of instruction. For participants who were given no warning ($M = 0.47$, $SD = 0.24$), primed lures elicited significantly more false insights than the other anagrams ($M = 0.02$, $SD = 0.04$), $t(252) = 17.26$, $p < .001$, CI = 0.38, 0.53. Consistent with our predictions, for participants who were given only a warning, primed lures ($M = 0.39$, $SD = 0.27$) produced significantly more false insights than control anagrams ($M = 0.02$, $SD = 0.05$), although this effect was slightly smaller, $t(252) = 13.97$, $p < .001$, CI = 0.29, 0.44. Unexpectedly, for those who were given a warning and an explanation of the task, primed lures ($M = 0.31$, $SD = 0.25$) still elicited more false insights than control anagrams ($M = 0.03$, $SD = 0.05$), $t(252) = 10.26$, $p < .001$, CI = 0.20, 0.36.

Exploratory analyses

We also examined whether the intensity ratings—originally taken to keep the FIAT protocol as similar to the original paradigm (Grimmer et al., 2022a) as possible—would also be affected by warnings. Perhaps the warnings reduced the intensity of the "Aha!" experience. To investigate this question, we conducted a 2 (Anagram Type: Primed Lure, Other) x 3 (Instruction Condition: Control, Warning, Warning + Explanation) mixed ANOVA on mean intensity ratings given to insights. This analysis revealed no significant main effect of Anagram Type, $F(1,235) = 0.10$, $p = .757$, $\eta^2_G < .001$, no main effect of Instruction Condition, $F(2,235) = 1.04$, $p = .356$, $\eta^2_G = .008$, and no interaction between Anagram Type and Instruction Condition, $F(2,235) = 0.21$, $p = .807$, $\eta^2_G < .001$.

Discussion

Can we prevent false insights as easily as we can create them? We tested whether warning participants that they would be lured into solving anagrams incorrectly could reduce the incidence of false insights. We found that warnings alone did not significantly reduce false insights. However, warning participants and fully explaining how the false insights were being elicited reduced—to a small degree—the false insight effect. These findings suggest that the false insight effect produced by the FIAT is a robust one, indicating that awareness of the processes that lead to the cognitive error can weaken but not prevent the effect. Or put differently, the illusion of insight generated under certain circumstances is a highly automatic process that is not easily overcome through conscious control.l

False insights versus false memories

Why did warnings not completely eliminate false insights? One reason is that the visual similarity manipulation may have given the paradigm some of the same qualities as visual illusions, which are notoriously difficult to overcome. Nevertheless, our findings were generally consistent with earlier studies using the DRM paradigm which found that warnings could weaken but not prevent the bias (Beato & Cadavid, 2015; Gallo et al., 1997; McDermott & Roediger, 1998; Peters et al., 2008; Yang et al., 2015). There are two main ways that the warnings could have diminished the effect of the FIAT. Either they operated at the level of encoding, making participants process the stimuli differently from the outset, or they acted upon participants' interpretation of their feelings of insight, making them wary of reporting their insight phenomenology.

As previous studies suggest, warnings may lead participants to process the task stimuli differently, as they are more effective in preventing false memories and illusory truth effects when presented prior to encoding rather than prior to responses (Jalbert et al., 2020; McCabe & Smith, 2002). On the other hand, the warnings may have caused participants to interpret their feelings of insight more cautiously, changing the metacognitive appraisal of their insight moments or making them more conservative in reporting them. It is not immediately clear which of these mechanisms warnings act upon. On the one hand, the intensity of insights did not change across conditions, which may indicate that warnings did *not* influence encoding since the presumably automatic experience of insight did not change. On the other hand, differences in encoding cannot be directly measured simply via intensity. For example, there may be a change in the *expected validity* of insight intensity at encoding (i.e. it is experienced but ignored). It is also challenging to test the metacognitive appraisal interpretation with the present data, because we do not know how metacognitive appraisal ought to affect (or not affect) insights and their intensity. Thus, at this stage, we cannot conclude whether warnings act at the level of encoding or metacognitive appraisal. Nonetheless, future research could investigate these mechanisms by asking participants to report on their approach to the task with additional self-report measures. For example, after each trial subjects could report on confidence, metacognitive reflection, and any overriding of initial responses.

Limitations and future directions

Some limitations are worth noting. One possibility is that the warnings might have had an immediate effect, but the effect quickly eroded as people focused on the task and forgot about the warnings. This experiment was not designed to test such a possibility, as its relatively short duration of only 12 trials does not allow for sufficient parsing by time, but future research could examine whether false insights are less common on trials immediately following the warning compared to later trials (or indeed include reminders at different stages of the trial). It is also possible that participants may have found it harder to avoid false insights due to the time restrictions, which may have prevented analytic reflection.

We also could not guarantee that all participants were sufficiently engaged in trying to avoid false insights. Perhaps if we had provided some form of incentive for correct responses, we may have found warnings to be more effective in reducing false insights. Aside from the comprehension check at the outset of the experiment, we did not measure the extent to which participants tried to follow the warnings.

It is important to consider the size of the effects we found, and what this might mean for future work on more applied questions. In our experiments, no matter the intervention—even when we gave a highly detailed description of how we were deceiving the participants—participants still experienced many false insights. Therefore, the key finding is the robustness of the false insight effect—primed lures reliably elicit false insights despite our best attempt to prevent them from doing so. Of course, the real question is whether false insights outside the lab are also difficult to prevent.

Although our task is based on solving anagrams, there are situations in everyday life that mimic the circumstances of the experiment. For example, often we are led towards a particular interpretation of events based on a carefully constructed narrative. A quintessential example is propaganda, wherein misleading or false information is selectively presented to induce a particular interpretation of events that subsequently unfold (such as a war). Thus, in the same way that the list of semantically associated words leads participants to a false "Aha!" regarding an ambiguous anagram, propaganda may similarly create the conditions for false insights by encouraging a particular interpretation of ambiguous events.

If false insights can be elicited in the real world, our findings raise further concern for how to prevent them with warnings. Further research is needed to better understand the mechanisms underlying false insights and how they are impacted by warnings. To elaborate on our current findings, future studies could test whether warnings can reduce false insights when presented after encoding but before the test (cf. Jalbert et al., 2020; McCabe & Smith, 2002). This manipulation would clarify whether warnings act upon metacognitive appraisal or task approach (i.e. encoding) and may help us understand why more detailed warnings reduce false insights more than simple ones. If we are correct to predict that warnings operate at the level of encoding, we would expect to see that warnings do not reduce false insights when presented before the test. The above test could also be complimented with the presence or absence of incentives to test whether the level of effort to avoid false insights improves the odds of reducing them.

Conclusion

In this study, we aimed to discover whether false insights could be reduced by warnings and explanations about the FIAT effect used to elicit them. We found that warning participants that they may be tricked into producing false solutions was not enough to reduce the false insight effect. Only when warnings were accompanied by a detailed explanation of how the FIAT paradigm works did participants experience fewer false insights. Nonetheless, participants who were given both a warning and an explanation were not completely inoculated against the FIAT effect, as they still reported a considerable number of false insights despite being aware of the conditions that produce them. These data suggest that if the conditions are ripe for false insights, then they are robust and difficult to prevent.

Notes

1. In our preregistration, there is an inconsistency in the strictness of this prediction between the information in our preregistration and the project wiki. As such, we present the weaker prediction that false insights will be reduced but not eliminated depending on the level of detail in the warnings.
2. We included these measures to keep the FIAT protocol as similar to the original study as possible. We did not analyse the memorie scores, however, as false memories were uncorrelated with false insights in (Grimmer et al., 2022a).

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Ruben E. Laukkonen http://orcid.org/0000-0001-8848-9231

References

Anastasi, J. S., Rhodes, M. G., & Burns, M. C. (2000). Distinguishing between memory illusions and actual memories using phenomenological measurements and explicit warnings. *The American Journal of Psychology*, *113*(1), 1–26. https://doi.org/10.2307/1423458

Aßfalg, A., & Bernstein, D. M. (2012). Puzzles produce strangers: A puzzling result for revelation-effect theories. *Journal of Memory and Language*, *67*(1), 86–92. https://doi.org/10.1016/j.jml.2011.12.011

Aßfalg, A., & Nadarevic, L. (2015). A word of warning: Instructions and feedback cannot prevent the revelation effect. *Consciousness and Cognition*, *34*, 75–86. https://doi.org/10.1016/j.concog.2015.03.016

Beato, M. S., & Cadavid, S. (2015). Normative study of theme identifiability: Instructions with and without explanation of the false memory effect. *Behavior Research Methods*, *48*(4), 1252–1265. https://doi.org/10.3758/s13428-015-0652-6

Bernstein, D. M., Whittlesea, B. W. A., & Loftus, E. F. (2002). Increasing confidence in remote autobiographical memory and general knowledge: Extensions of the revelation effect. *Memory & Cognition*, *30*(3), 432–438. https://doi.org/10.3758/BF03194943

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

Calvillo, D. P., & Parong, J. A. (2016). The misinformation effect is unrelated to the DRM effect with and without a DRM warning. *Memory*, *24*(3), 324–333. https://doi.org/10.1080/09658211.2015.1005633

Gallo, D. A., Roberts, M. J., & Seamon, J. G. (1997). Remembering words not presented in lists: Can we avoid creating false memories? *Psychonomic Bulletin & Review*, *4*(2), 271–276. https://doi.org/10.3758/BF03209405

Grimmer, H. J., Laukkonen, R. E., Tangen, J. M., & von Hippel, W. (2022a). Eliciting false insights with semantic priming. *Psychonomic Bulletin & Review*, *29*, 954–970. https://doi.org/10.3758/s13423-021-02049-x

Grimmer, H. J., Tangen, J. M., Laukkonen, R. E., von Hippel, W., & Freydenzon, A. (2022b). Thinking style and psychosis proneness do not predict false insights. https://doi.org/10.31234/osf.io/uctn2.

Jalbert, M., Newman, E., & Schwarz, N. (2020). Only half of what I'll tell you is true: Expecting to encounter falsehoods reduces illusory truth. *Journal of Applied Research in Memory and Cognition*, *9*(4), 602–613. https://doi.org/10.1016/j.jarmac.2020.08.010

Kounios, J., & Beeman, M. (2009). The Aha! moment: The cognitive neuroscience of insight. *Current Directions in Psychological Science*, *18*(4), 210–216. https://doi.org/10.1111/j.1467-8721.2009.01638.x

Kronlund, A., & Bernstein, D. M. (2006). Unscrambling words increases brand name recognition and preference. *Applied Cognitive Psychology*, *20*(5), 681–687. https://doi.org/10.1002/acp.1220

Lakens, D., Scheel, A. M., & Isager, P. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Laukkonen, R. E., 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

McCabe, D. P., & Smith, A. D. (2002). The effect of warnings on false memories in young and older adults. *Memory & Cognition*, *30*(7), 1065–1077. https://doi.org/10.3758/BF03194324

McDermott, K. B., & Roediger, H. L. (1998). Attempting to avoid illusory memories: Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. *Journal of Memory and Language*, *39*(3), 508–520. https://doi.org/10.1006/jmla.1998.2582

Nadarevic, L., & Aßfalg, A. (2017). Unveiling the truth: Warnings reduce the repetition-based truth effect. *Psychological Research*, *81*(4), 814–826. https://doi.org/10.1007/s00426-016-0777-y

Newman, E. J., Jalbert, M. C., Schwarz, N., & Ly, D. P. (2020). Truthiness, the illusory truth effect, and the role of need for cognition. *Consciousness and Cognition*, *78*, 102866. https://doi.org/10.1016/j.concog.2019.102866

Nichols, R. M., & Loftus, E. F. (2019). Who is susceptible in three false memory tasks? *Memory*, *27*(7), 962–984. https://doi.org/10.1080/09658211.2019.1611862

Otgaar, H., Howe, M. L., Muris, P., & Merckelbach, H. (2019). Associative activation as a mechanism underlying false memory formation. *Clinical Psychological Science*, *7*(2), 191–195. https://doi.org/10.1177/2167702618807189

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Salvi, C., Bricolo, E., Kounios, J., Bowden, E., & 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

Starns, J. J., Lane, S. M., Alonzo, J. D., & Roussel, C. C. (2007). Metamnemonic control over the discriminability of memory evidence: A signal detection analysis of warning effects in the associative list paradigm. *Journal of Memory and Language*, *56*(4), 592–607. https://doi.org/10.1016/j.jml.2006.08.013

Peters, M. J. V., Jelicic, M., Gorski, B., Sijstermans, K., Gisbrecht, T., & Merckelbach, H. (2008). The corrective effects of warning on false memories in the DRM paradigm are limited to full attention conditions. *Acta Psychologica*, *129*(2), 308–314. https://doi.org/10.1016/j.actpsy.2008.08.007

Unkelbach, C., & Greifeneder, R. (2018). Experiential fluency and declarative advice jointly inform judgements of truth. *Journal* *of Experimental Social Psychology*, *79*, 78–86. https://doi.org/10.1016/j.jesp.2018.06.010

Watson, J. M., Bunting, M. F., Poole, B. J., & Conway, A. R. (2005). Individual differences in susceptibility to false memory in the Deese-Roediger-McDermott paradigm. *Journal Of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 76–85. https://doi.org/10.1037/0278-7393.31.1.76

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*, *55*(2), 289–293. https://doi.org/10.1002/jocb.433

Yang, H., Yang, S., Ceci, S. J., & Isen, A. M. (2015). Positive affect facilitates the effect of a warning on false memory in the DRM paradigm. *The Journal of Positive Psychology*, *10*(3), 196–206. https://doi.org/10.1080/17439760.2014.950177

Zhu, B., Chen, C., Loftus, E. F., Lin, C., & Dong, Q. (2013). The relationship between DRM and misinformation false memories. *Memory & Cognition*, *41*(6), 832–838. https://doi.org/10.3758/s13421-013-0300-2

Appendices

Appendix A: instructions video transcripts

Control Condition.

In this experiment, you'll begin by studying a list of words. Your job is to remember as many of these words as possible. Let's go through an example. Are you ready? HYGIENE, CLEANLINESS, SEWERAGE, DISINFECTION, ANTISEPTIC, TOILETS, WASTE, HEALTH, CONTAMINATION, WASHING. Good! Now we'll come back to this list of words in a minute. But first, you're going to solve a scrambled word problem. You'll see a scrambled word like this: SANIIOTTRN. Your job is to unscramble it into an English word as quickly as you can. Once you think you know what the word is during the experiment, you'll click on the arrow button. Can you solve the scrambled word above?

Good! The correct answer is … TRANSITION. Let's try another one. Ready?: INTSAEPTIC. Good! The correct answer is … ANTISEPTIC.

So, you'll be given a minute or so to unscramble the word. If you haven't solved it during that time, we'll ask you to make a guess.

Once you've typed your response we'll ask you whether you had a moment of insight or not. When you were first presented with the scrambled letters, they didn't make any sense in that order. But at some point, this unsolvable puzzle may have suddenly become clear and obvious. This sudden and unexpected solution is what we mean by a moment on insight. Think of it as a miniature Eureka or lightbulb moment. We'll start by asking you whether you had an insight moment or not. And if you click "yes" then we'll ask you to rate the intensity of this insight moment on a scale from 1 to 10. A rating of 1 on this scale means you had a very weak insight, and a rating of 10 mean that you had a very strong insight. After you make your rating, we'll ask you to recall as many words from the original study list as you can. You'll do this by typing as many words as you can remember.

When you've listed as many words as you can, click the arrow button to continue. And we'll present you with another list of words to remember, and repeat this process a few more times.

Next, we're going to show you a practice trial of the task so you'll be ready to do the actual experiment. Remember: you'll see a list of words, solve some anagrams, tell us whether you had an insight moment … If so, then you'll rate its intensity, and then recall as many words as you can. That's all there is to it!

Warning Condition

You'll begin by studying a list of words. Your job is to remember as many of these words as possible. Let's go through an example. Are you ready? HYGIENE, CLEANLINESS, SEWERAGE, DISINFECTION, ANTISEPTIC, TOILETS, WASTE, HEALTH, CONTAMINATION, WASHING. Good! Now we'll come back to this list of words in a minute. But first, you're going to solve a scrambled word problem. You'll see a scrambled word like this: SANIIOTTRN. Your job is to unscramble it into an English word as quickly as you can. Once you think you know what the word is during the experiment, you'll click on the arrow button. Can you solve the scrambled word above?

Good! The correct answer is … TRANSITION. Let's try another one. Ready?: INTSAEPTIC. Good! The correct answer is … ANTISEPTIC.

So, you'll be given a minute or so to unscramble the word. If you haven't solved it during that time, we'll ask you to make a guess.

Once you've typed your response we'll ask you whether you had a moment of insight or not. When you were first presented with the scrambled letters, they didn't make any sense in that order. But at some point, this unsolvable puzzle may have suddenly become clear and obvious. This sudden and unexpected solution is what we mean by a moment on insight. Think of it as a miniature Eureka or lightbulb moment. We'll start by asking you whether you had an insight moment or not. And if you click "yes" then we'll ask you to rate the intensity of this insight moment on a scale from 1 to 10. A rating of 1 on this scale means you had a very weak insight, and a rating of 10 mean that you had a very strong insight. After you make your rating, we'll ask you to recall as many words from the original study list as you can. You'll do this by typing as many words as you can remember.

When you've listed as many words as you can, click the arrow button to continue. And we'll present you with another list of words to remember, and repeat this process a few more times.

Beware: half of the anagrams have been designed to trick you into giving the wrong answer. Watch out for these anagrams and do your best to avoid being lures into giving the wrong answer!

Next, we're going to show you a practice trial of the task so you'll be ready to do the actual experiment. Remember: you'll see a list of words, solve some anagrams, tell us whether you had an insight moment … If so, then you'll rate its intensity, and then recall as many words as you can. And beware of the trick anagrams! That's all there is to it!

Explanation Condition

In this experiment, you're going to complete a False Insight Task. This is a tricky task that's designed to elicit feelings of insight for incorrect solutions. We'll do this by having you study a list of words that will lead you to think of a certain concept. For example, hill, valley, climb, summit, top, molehill, peak, goat, steep, ski. And then we'll present you with a scrambled word and ask you to unscramble it: MOUNIAIMNT. In this case, the solution that might have popped into your head might have been … MOUNTAIN. But, this would be incorrect, because there is an extra "M" and an extra "I" in this solution. In fact the correct solution is AMMUNITION. But the list of words we presented was designed to lure you into linking of MOUNTAIN … And the configuration of the letters in the anagram looked fairly similar to MOUNTAIN … So you might have incorrectly felt like you'd come up with the right solution. This is what we mean by a false insight, and this is what we're investigating in this experiment … Here's what will happen during the experiment … You'll begin by studying a list of words. Your job is to remember as many of these words as possible. Let's go through an example. Are you ready? HYGIENE, CLEANLINESS, SEWERAGE, DISINFECTION, ANTISEPTIC, TOILETS, WASTE, HEALTH, CONTAMINATION, WASHING. Good! Now we'll come back to this list of words in a minute. But first, you're going to solve a scrambled word problem. You'll see a scrambled word like this: SANIIOTTRN. Your job is to unscramble it into an English word as quickly as you can. Once you think you know what the word is during the experiment, you'll click on the arrow button. Can you solve the scrambled word above? Now remember that this is a false insight task, so given the list of words you studied and the look of the scrambled word above, the word SANITATION might come to mind. But this is incorrect. The correct answer is TRANSITION. Let's try another one. Ready?: INTSAEPTIC. Not all of the scrambled words you'll see are false, so you have to be careful … In this case the correct answer is ANTISEPTIC. So, you'll be given a minute or so to unscramble the word. If you haven't solved it during that time, we'll ask you to make a guess.

Once you've typed your response we'll ask you whether you had a moment of insight or not. When you were first presented with the scrambled letters, they didn't make any sense in that order. But at some point, this unsolvable puzzle may have suddenly become clear and obvious. This sudden and unexpected solution is what we mean by a moment on insight. Think of it as a miniature Eureka or lightbulb moment. We'll start by asking you whether you had an insight moment or not. And if you click "yes" then we'll ask you to rate the intensity of this insight moment on a scale from 1 to 10. A rating of 1 on this scale means you had a very weak insight, and a rating of 10 mean that you had a very strong insight. After you make your rating, we'll ask you to recall as many words from the original study list as you can. You'll do this by typing as many words as you can remember.

When you've listed as many words as you can, click the arrow button to continue. And we'll present you with another list of words to remember, and repeat this process a few more times.

Beware: half of the anagrams have been designed to trick you into giving the wrong answer. Watch out for these anagrams and do your best to avoid being lures into giving the wrong answer!

Next, we're going to show you a practice trial of the task so you'll be ready to do the actual experiment. Remember: you'll see a list of words, solve some anagrams, tell us whether you had an insight moment … If so, then you'll rate its intensity, and then recall as many words as you can. And beware of the trick anagrams! That's all there is to it!

Appendix B: comprehension check questions

Correct answers are in bold.

Control Condition

"In the task I am about to complete, my most important job is to:"

a) Solve the anagrams correctly
b) **Remember the words and solve anagrams as quickly as possible**
c) Remember the words correctly

Warning Condition

"During the instructions, you were given a warning to avoid:"

a) Being lured into solving the anagrams incorrectly
b) Being tricked into forgetting the study words
c) Being tricked into spending too much time on the task

Warning + Explanation Condition

"During the instructions, you learned that the task contains a trick to lure you into giving the wrong answer. What is the trick?"

a) Some scramble words will be impossible to solve
b) Some scrambled words will not be in English
c) **Some scrambled words will look like words that are related to those you studied**