Journal of Problem Solving

SPECIAL ISSUE

Problem Solving as an Encoding Task: A Special Case of the Generation Effect

Jasmin M. Kizilirmak,[1] Berit Wiegmann,[2] and Alan Richardson-Klavehn[2]

[1] University of Hildesheim, [2] Otto-von-Guericke-University of Magdeburg, Germany

Correspondence:
Correspondence concerning this article should be addressed to Jasmin M. Kizilirmak, via email to kizilirmak@uni-hildesheim.de.

Keywords:
insight, problem solving, generation effect, long-term memory, learning

Acknowledgment:
This work was supported by a grant assigned to A. R.-K. by the German Research Foundation (Deutsche Forschungsgemeinschaft) for project TPA10N, part of the collaborative research center "Neurobiology of Motivated Behavior" SFB779, awarded to the University of Magdeburg.

Recent evidence suggests that solving problems through insight can enhance long-term memory for the problem and its solution. Previous findings have shown that generation of the solution as well as experiencing a feeling of Aha! can have a beneficial relationship to later memory. These findings lead to the question of how learning in problem-solving tasks in which a novel solution needs to be generated—such as in tasks used to study insight—differs from the classical generation effect. Because previous studies on learning from insight on one hand and the generation effect on the other hand have measured different types of memory, the present study examined two kinds of memory measures: indirect (solving old and new problems at test) and direct (recognition memory). At encoding, we manipulated whether participants had the chance to solve Compound Remote Associates Task items and compared later memory for generated solutions (generate condition) to solutions that were presented after failing to generate one (fail-to-generate condition), and to solutions that were presented without a chance at generation (read condition). Participants also reported if they had an Aha! experience for each problem. While both Aha! experiences and generated solutions were associated with more positive emotional responses, only the generation variable was associated with differences in later memory performance. While attempts to generate had an advantage over the read condition in recognition memory performance (generate > fail-to-generate > read), only when generation was successful did it enhance the solution rate of old items during testing (generate > read > fail-to-generate). Contrary to generation effects with other verbal stimuli, these results suggest that the generation effect in problem-solving tasks in which a novel solution needs to be found differs from the classical generation effect. Seeing a correct solution for a longer time (read) seems in the current case to be more helpful for solving the same problem later on, compared to being presented with the solution after a failed attempt at problem solving.

Previous findings suggest that solving problems through insight can enhance later memory for the problem and solution (Ash, Jee, & Wiley, 2012; Danek, Fraps, von Müller, Grothe, & Öllinger, 2013a; Dominowski & Buyer, 2000; Kizilirmak, Galvao Gomes da Silva, Imamoglu, & Richardson-Klavehn, 2015). The core definition of an insight, as considered in this paper, is the sudden comprehension of a relationship that until that moment appeared incomprehensible. A famous example of such a sudden insight is the anecdote of Archimedes's discovery of the relationship between the volume submerged and the water level rising by stepping into the bathtub. In a previous study employing a pictorial problem-solving task, successfully generating a solution as compared to being presented with a solution after failing to generate one, experiencing insight (i.e., a feeling of Aha! consisting of surprise and being convinced of the truth of the solution) as compared to no insight, and the positive emotional response to successful generation and insight, were all beneficially related to learning (Kizilirmak et al., 2015). Here, we further illuminate the impact of generating a solution to a problem by additionally examining a "no-chance-to-solve" condition in which the solution was immediately provided together with the problem, thus bypassing the chance to generate, thereby linking our study more closely with the numerous traditional memory studies on the generation effect (see next subsection). Moreover, a verbal problem-solving task is used, the Compound Remote Associates Task (CRAT), because verbal stimulus material has usually been used to study the generation effect. The CRAT has been utilized to investigate insightful problem solving in other studies (e.g., Bowden & Jung-Beeman, 2003; Cranford & Moss, 2010). In the current study, the subjective Aha! experience and performance on the CRAT (i.e., generating versus not generating the correct solution to a problem) are both examined in relation to later memory for solutions.

http://dx.doi.org/10.7771/1932-6246.1182

THE GENERATION EFFECT

While the subjective Aha! experience may be unique to insight, the generation of a solution and its positive relationship to later memory are not. The beneficial effect of insight on learning is therefore likely partially based on the positive effect of generation on long-term memory formation. The so-called generation effect describes the superiority of generated items (generate condition) over presented items (read condition) regarding later long-term memory performance (Burns, 1992; Slamecka & Graf, 1978). The generate condition typically involves the production of a word, such as coming up with a semantic associate or completing a word fragment. The processes behind the beneficial effect of generation on memory encoding are not completely understood, but it is generally agreed that better semantic integration plays an important role (McElroy & Slamecka, 1982; Slamecka & Graf, 1978). Higher cognitive effort and a deeper level of processing have also been considered, but insufficient evidence has been found to support those claims (Bertsch, Pesta, Wiscott, & McDaniel, 2007). The key question arises as to whether the generation effect in problem solving is a special case of the generation effect or just a typical case of the generation effect.

The generation effect is investigated by comparing a generate condition in which the target item has to be produced by the participant, with a read condition, in which the target item does not need to be generated by the participant but is instead presented. The read condition does not necessarily imply verbal item material. It has only become convention to call the condition in which the stimulus is directly presented "read," because early generation effect studies used verbal material. The generation rule is typically easy enough so that the vast majority of the items can be successfully generated, in order to avoid an item selection effect, that is, where only easy items are successfully generated (Gardiner, Java, & Richardson-Klavehn, 1996; Slamecka & Graf, 1978).[1] For most problem-solving tasks that are used to investigate insight, this is not the case. These types of items require creative thinking and are often made to be almost unsolvable via normal analytical problem-solving strategies (e.g., Mednick, 1962). Typically, a representational change is required to solve the item, that is, to look at the problem and solution space from a different angle (Knoblich, Ohlsson, Haider, & Rhenius, 1999). Hence, especially with the time limit required in laboratory experiments, and with some individuals being more creative than others, it cannot be avoided that a considerable number of items are not going to be solved. In problem-solving tasks used to investigate insight one will therefore not only have a generate and read condition, but also most certainly a considerable number of fail-to-generate trials.

One study on the classical generation effect also considered a fail-to-generate condition and showed that later memory is still higher for fail-to-generate than for read, given that the correct response is provided after a failed attempt to generate (Slamecka & Fevreiski, 1983). Slamecka and Fevreiski found that when people were asked to generate the opposites for words, fail-to-generate enhanced free recall performance (compared to read) 24 hours later as much as did generate (generate = fail-to-generate > read). In comparison, fail-to-generate did not enhance recognition memory performance as much as generate, although it was still better than read (generate > fail-to-generate > read). The authors proposed that a word's semantic-associative attributes have to be activated for subsequent recall to be successful, which may also be the case for fail-to-generate, which was proposed to represent incomplete generation. Incomplete generation was thought of as the generation of semantic attributes of the solution without arriving at the proper lexical entry. Therefore, recognition memory, which relies on both surface and semantic properties of the word, benefits less from fail-to-generate than does free recall. Whether fail-to-generate can also be considered to represent incomplete generation in problem-solving tasks that have mainly been used to study insight, is an important question that may dissociate generation effects on this type of problem from the classical generation effect. Considering that first ideas about the solution are often misleading and that problem-solving attempts for CRAT problems frequently lead to an impasse that needs to be overcome (Bowden & Jung-Beeman, 1998; Chein & Weisberg, 2014; Cranford & Moss, 2012; Hoffman & Subramaniam, 1995), the fail-to-generate condition in a CRAT problem-solving paradigm may not benefit from incomplete generation.

THE AHA! EXPERIENCE AND LATER MEMORY

As previously mentioned, the occurrence of the subjective Aha! experience is often associated with performance in certain problem-solving tasks, such as the CRAT. Recent studies on insight have assessed the experience of such an Aha! moment by asking the participants about it, and have also shown that later memory performance for problems encoded with Aha! is better than for problems encoded without Aha! (Danek et al., 2013a; Dominowski & Buyer, 2000; Kizilirmak et al., 2015). Such an Aha! experience is typically defined as the comprehension of the solution being sudden and unexpected—that participants are completely convinced of the correctness of the solution, and that the solution is in hindsight being experienced as easy and very clear (Danek, Fraps, von Müller, Grothe, & Öllinger, 2014; Topolinski & Reber, 2010). Also, the Aha! experience is often accompanied by a positive feeling. Previous studies that investigated insight and learning have focused on the re-solution effect, in other words the ability to solve problems which have been solved via an insight during an encoding phase, and again during a testing phase (Ash et al., 2012; Dominowski &

Buyer, 2000). It has been reported that solution times were significantly reduced and solution rates were significantly higher for previously solved problems. However, re-solving problems is only one way to assess learning success. More importantly, generation effect studies have generally used a different type of memory measure, as explained under the subsection "Direct Versus Indirect Memory Tests."

In most studies in which the Aha! experience has been assessed, this is only done for generated solutions. Generated solutions with Aha! are then designated as "insight" solutions. However, Kizilirmak et al. (2015) asked participants whether they experienced an Aha! moment not only when they generated a solution, but also when they failed to do so and were shown the solution instead. The definition of an Aha! experience was phrased such that generation was not a necessary precondition; the solution could still be comprehended suddenly, appearing utterly convincing and clear as day when it was comprehended after it was presented following a failed attempt at problem solving. Using this procedure, Kizilirmak et al. (2015) found that the Aha! experience was accompanied by a more positive feeling and better memory performance independent of whether the solution to a problem was generated or not. Thus, in the current study participants were asked whether they experienced an Aha! both when they successfully solved a problem on their own and when they did not. The main design difference from the prior study was that the addition of the equivalent of a read condition in the generation-effect studies, in which the solution was presented immediately, and the participant therefore had no chance to find the solution on their own.

DIRECT VERSUS INDIRECT MEMORY TESTS

There are different ways in which long-term memory can be tested, and as seen in Slamecka and Fevreiski (1983), different measures may be more affected by generation than others. One could classify the different types of tests as direct and indirect memory tests (see Richardson-Klavehn & Bjork, 1988, and Richardson-Klavehn, 2010, for extensive overviews). A test is considered direct/intentional when it is clearly stated in the instructions for the participant that memory content should be recalled/recognized. Such tests are thought to mainly tap into controlled, voluntary retrieval processes. On the other hand, a test is considered to be indirect or incidental when the task can in principle be solved based on current skills and knowledge without the necessity to make mental reference to previous experimental episodes. The instructions for such tests do not include the instruction to remember anything from a previous experimental learning episode; indeed, participants may even be explicitly instructed to ignore episodic memory. Such indirect memory tests are thought to mainly tap into automatic, involuntary retrieval processes, and it has even been found that they sometimes show evidence of memory even when participants lack consciousness of the prior episodes whose influence is being revealed in their performance.

Nevertheless, it is very important to make a distinction between the type of test (as defined objectively by instructions) and the type of retrieval process accessed by the test (see Richardson-Klavehn & Bjork, 1988). Although the instructions are those of an indirect or incidental test, the participants might not only use involuntary retrieval processes to solve the task. For example, consider an experiment in which participants try to solve puzzles in a first session (learning phase) and are tested on solving those same puzzles and new ones in a second session (memory test). The test instructions do not say that they should try to remember the solutions to those problems from the first session. While solving old problems, participants could involuntarily benefit from their previous encounter with the problems. However, it could also be that they consciously recognize an old problem and voluntarily recall the previous solution. It could even be that they first unconsciously use their knowledge from their previous encounter with an old problem until they realize this consciously and then voluntarily recall the previously learned information. Similarly, direct or intentional memory tests can also profit from involuntary memory. Thus, one can only say that direct and indirect memory tests predominantly tap into more voluntary versus more involuntary retrieval processes, respectively. There are ways to clearly dissociate voluntary and involuntary retrieval processes (Richardson-Klavehn & Bjork, 1988; Richardson-Klavehn, 2010; see Schott et al., 2005, for neural evidence), but such a dissociation was not attempted in the current study.

Most relevant here is that in the problem-solving literature, learning from insight has been mainly investigated with indirect or incidental memory tests. The memory benefit from repeatedly solving old items has been termed the *resolution effect* (Dominowski & Buyer, 2000). The generation effect, by contrast, has been investigated with many different types of memory tests, of which most can be considered direct or intentional memory tests. In the 86 studies analyzed in the meta-analysis by Bertsch et al. (2007), the memory test was either a recognition memory test (e.g., an item is presented, and an old or new decision has to be made), a cued-recall test (e.g., one word of a previously studied word pair is shown, and the other word has to be retrieved), or a free recall test (all studied stimuli have to be recalled in any order). The current study, therefore, employed both direct/intentional and indirect/incidental memory tests to gain insight into both relatively more voluntary as well as relatively more involuntary forms of memory for previously encountered CRAT problems. Doing so created a further bridge between problem-solving and memory studies.

AIMS OF THE CURRENT STUDY

The aim of the two experiments was to investigate the effects of having the chance to try to solve a CRAT problem, or not having the chance to solve the problem at all, with the solution being immediately presented, on long-term memory formation. Moreover, the chance-to-solve category was divided, post-hoc, into successful generation (note that incorrect generations were infrequent and not further analyzed) and failed generation (there was no solution at all within the time limit, but a solution was presented as feedback), and further split for all categories as to whether an Aha! was experienced or not. The studies employed a version of the Compound Remote Associates Task (CRAT) which was adapted from Bowden and Jung-Beeman (2003), whose task was based on the Remote Associates Task originally designed by Mednick (1962) to test creativity in students. In the CRAT, three nouns (i.e., a triad) are presented that have no obvious association and are therefore considered to be remote associates, for example: manners, cloth, and tennis. Close associates, such as tennis-racquet, to single triad words are even misleading (for the solution, see Figure 1 below). The task is to find a fourth word that can be used to build compound words with each of the other three, thereby building a connection between all four words. Two different measures of later memory were used to evaluate relationships to the encoding conditions just described: (1) more involuntary memory by means of an indirect memory test of solutions (Experiment 1), and (2) more voluntary memory by means of a direct memory test in which an old or new decision was performed either on the complete item (Experiment 1) or on the solution word alone (Experiment 2). The experiments were similar in the encoding phase, but differed in regard to the memory tests. Learning was incidental because participants were not informed that their memory would be tested later on. The indirect memory test (Experiment 1) was designed to primarily assess the involuntary retrieval of the solution upon the presentation of the triad, and therefore assessed the encoding of the association between problem and solution, whereas the intentional memory test primarily assessed episodic memory, either for the triad in combination with its solution (Experiment 1), or for the solution without the presence of the triad (Experiment 2). The indirect memory test in Experiment 1 was most comparable with Slamecka's and Fevreiski's free recall procedure, because both heavily rely on activation spread in associative memory networks (Collins & Loftus, 1975), whereas the direct memory test was comparable to Slamecka's and Fevreiski's recognition procedure.

HYPOTHESES

Participants were expected to engage in semantic processing in all conditions, because the connection between the four words needed to be understood in all of them. It was further hypothesized that the resulting association between the triad and solution would vary in degree. As the semantic-associative processing ends successfully in the generate condition, the association between the triad and solution should be strongly enhanced. However, in contrast to classical generation effect studies that employ tasks in which the solution may be found by analytical problem solving, with successive attempts gradually coming closer to the solution, thinking of close associations with the triad words in the CRAT is not supposed to lead to the solution. Thus, memory performance in the fail-to-generate condition for CRAT problems may differ from the benefit seen in the Slamecka and Fevreiski study (1983), because the semantic-associative processing does not directly lead closer to the solution. It was hypothesized that failing to generate the solution might even lead to interference between the correct solution and the incorrect solutions processed during failed problem solving (Kane & Anderson, 1978). As incorrect solutions are not considered by the participant in the read condition, the association between the correct solution and triad might be even stronger than in the fail-to-generate condition. For the indirect memory test that relies on associative memory, the following line-up of memory performance may be expected: generate > read > fail-to-generate. Alternatively, if searching for the solution leads to spreading activation that eventually reaches the correct remote association that is the solution word, which then suddenly emerges into consciousness (Bowers, Regehr, Balthazard, & Parker, 1990), the same pattern as found by Slamecka and Fevreiski (1983) could be expected: generate = fail-to-generate > read.

Regarding the direct memory test that relies on recognition memory, the hypothesis matched the order reported by Slamecka and Fevreiski (generate > fail-to-generate > read), because recognition, if mainly based on a feeling of familiarity, should not depend on the strength of the association between triad and solution. This pattern would be especially expected when recognition is tested for the solution word alone (Experiment 2). However, recollection does depend on remembering some details about the encoding episode (Gardiner, Ramponi, & Richardson-Klavehn, 1998), for which relational encoding is especially important (Yonelinas & Ritchey, 2015). Moreover, emotional arousal during stimulus encoding has specifically been reported to enhance recollection (Anderson, Yamaguchi, Grabski, & Lacka, 2006; Yonelinas & Ritchey, 2015). It would therefore be plausible that recollection will be better for more emotional encoding episodes (generate: feeling of success; fail-to-generate: frustration, disappointment). Thus, if recognition memory performance is mainly based on recollection, generate and fail-to-generate at encoding might result in better later memory than read (generate = fail-to-generate > read).

As mentioned earlier, in addition to generation, the subjective experience of insight during the comprehension of the solution on memory for solutions was also analyzed. This was done by asking participants whether they had an

Aha! experience and how happy or unhappy they felt during comprehension. The subjective Aha! experience has been investigated in insight studies before, and suggests a qualitative difference not only during the processing of the solution (Jung-Beeman et al., 2004; Kounios et al., 2006), but also in relation to later memory performance (Danek et al., 2013a; Kizilirmak et al., 2015). Based on previous findings (Danek et al., 2013a; Kizilirmak et al., 2015), Aha! experiences were expected to be associated with a relatively better memory performance, compared to solutions understood without Aha!

Because the Aha! experience is usually described to be accompanied by a positive feeling (Danek et al., 2013a; Jung-Beeman et al., 2004), it was hypothesized that the generate condition and Aha! experiences would be associated with a more positive emotional response than the fail-to-generate and read conditions, and no Aha! experiences, with fail-to-generate possibly evoking a more negative emotional response than read. Such a finding would validate the intrinsically rewarding aspect of the Aha! experience.

EXPERIMENT 1

The main aim of Experiment 1 was to test the relationship between (1) generate, fail-to-generate, and read solutions, and (2) the subjective feeling of Aha! or its absence, with later memory performance measured with both an indirect and a direct test after one week. The indirect test measured the solution rate of old items, in comparison to the solution rate of new items. The direct test measured the rate of items (their solution included) correctly recognized as old. These tests were combined to form a hybrid test (see Richardson-Klavehn, 2010; Schott et al., 2005), with an indirect and then a direct test on each test item.

METHODS

Participants

Participants were 21 German native speakers (7 male, 14 female) with a mean age of 24.1 years (SD = 2.4, range: 20–28). They participated after giving written informed consent. Participants were informed that they had the right to abort the experiment at any time without any negative personal consequences. Participation was paid with 6 Euros per hour. At the end of the experiment, participants were debriefed on request about the purpose of the study. Five participants were excluded from statistical analysis of the GENERATION (generate, fail-to-generate, read) × AHA (Aha!, no Aha!) fully crossed design, because of empty cells in some conditions. The remaining 16 participants (4 male, 12 female) had a mean age of 23.7 years and were all university students or had a university degree.

Stimulus material and apparatus

A German adaptation of the CRAT (Bowden & Jung-Beeman, 2003) with 180 items was used for the experiment. All items consisted of nouns or color words. The solution could either be used as a prefix or a suffix with each of the triad words to form a compound word. Table 1 shows three examples. Some items were homogeneous, i.e., the solution could be affixed to all of the triad words in the same way (only as a prefix or only as a suffix, see Example 1, Table 1), while others were heterogeneous, that is, whether the solution could be used as a prefix or suffix varied between the triad words (Example 2, Table 1). Since the authors of the original task found no performance difference between heterogeneous and homogeneous items, these were kept approximately equal. Due to the nature of the German language, in some cases one word needed to be slightly modified to form a compound word with another (see Example 3, Table 1, "Meile" and "Stein" can be combined to "Meilenstein"). Solution words were never repeated or used as triad words. Triad words were rarely repeated but not more than twice.

On the basis of unpublished normative data, the lists were split into four sublists that matched regarding (1) the solution rate, (2) plausibility, and (3) the probability of an Aha! experience accompanying comprehension of the solution. Two lists were assigned to the chance-to-generate condition (90 items, later split into generate and fail-to-generate based on participants' responses), one list to the no-chance-to-generate (read) condition in which the solution was

Table 1
Examples of Compound Remote Associate items.

Triad words	Solution word	Compund word
(1) Manieren, Tennis, Tuch	Tisch	Tischmanieren, Tischtennis, Tischtuch
manners, tennis, cloth	table	table manners, table tennis, table
(2) Stufe, System, Feuer	Alarm	Alarmstufe, Alarmsystem, Feueralarm
level, system, fire	alarm	alarm level, alarm system, fire alarm
(3) Kiesel, Meile, Zeit	Stein	Kieselstein, Meilenstein, Steinzeit
pebble, mile, age	stone	pebble stone, mile stone, stone age

immediately presented (45 items), and one list was used as new items in the memory test (45 items). The assignment of lists to conditions was counterbalanced across participants by means of a reduced Latin square. However, as some participants were excluded from data analysis, the assignment of lists to participants employed a compromise of following the Latin square while still acquiring enough participants per condition for the analysis.

The Aha! or no Aha! decision was made on the basis of a written definition of the subjective Aha! experience, which was provided as part of the instruction for the learning phase. The definition of the Aha! experience read approximately as follows (translated from German) and was based on the criteria listed by Topolinski and Reber (2010), with the exception of the "positive emotional response" criterion because this was evaluated separately:

> By 'aha!' experience we are referring to the feeling of a sudden insight, that is, the surprising comprehension for a previously seemingly unsolvable problem. At this moment of insight, you are convinced of the correctness of the solution. The solution suddenly appears to be as plain as day. If you find the solution on your own, you are usually unaware of how it came into your mind. The described feeling of 'aha!' does not have to be overwhelming, but should closely correspond to this description.

The emotional response to comprehending the solution was measured by means of a 5-point graphical smiley interval scale (numerical values: -2 to +2) ranging from very sad (mouth: upper half of a circle, -2) to neutral (mouth: flat horizontal line, 0) to very happy (mouth: lower half of a circle, +2).

A standard desktop PC with Windows XP (Microsoft, Redmond Campus, Washington, USA) was used and stimuli were presented on a 19" TFT screen with a 60 Hz frame rate and 1280 × 1024 pixels resolution. Stimulus presentation and collection of behavioral data were controlled with Presentation 16.3 (Neurobehavioral Systems, Berkeley, CA). All responses were made with a standard USB keyboard. Statistics were analyzed with SPSS 22 and 23 for Windows 7 64-bit (IBM, Armonk, NY, USA) and the effect size calculator (used to calculate Cohen's *d* provided by Melody Wiseheart on http://www.cognitiveflexibility.org/effectsize).

Design

The design was a 3 × 2 within-subjects design. Of main interest was the relationship between GENERATION (generated, fail-to-generate, read), AHA (Aha!, no Aha!), and later memory performance (either the solution rate of old items, or the recognition rate of old items) and on the emotional response during the comprehension of the solution. A priori, two conditions formed an experimentally manipulated variable: items with a chance to generate a solution, whereby the solution was either successfully generated or was, in the case of generation failure, provided after a time limit ran out, and items with an immediately provided solution and thus no chance to generate (read condition). The former were post-hoc split into generate (correctly solved) and fail-to-generate (no solution within time limit). Incorrect solutions were discarded from analysis. Memory performance was tested in two ways: first, with an indirect memory test, in which subjects were asked to try solving old and new problems regardless of whether they were perceived as old or new, followed by a direct memory test, that is, an old or new recognition memory task.

To avoid the perceived plausibility of any provided solution complicating the results due to, for example, individual differences in vocabulary, we asked participants to make a plausible or implausible judgment for each item (see Figures 1 and 2) and excluded all items rated as implausible in both the learning and/or test phase from further analysis. Moreover, generated solutions were manually rated as correct or incorrect by the experimenter before data analysis (as in the original RAT by Mednick, 1962, all items supposedly only had one possible correct solution). When participants solved a problem incorrectly, it could have a number of complicated effects on remembering the correct solution later on (e.g., cognitive interference between the correct and incorrect solutions during encoding/retrieval). Such incorrectly solved trials were rare and were therefore excluded from all statistical analyses.

Task and procedure

The experiment consisted of two sessions: a learning phase and a testing phase that was conducted one week after the learning phase. In the learning phase, participants were presented with CRA items and were either instructed to try to solve the items or just to try to understand the solution when it was provided. Since it has been reported that the generation effect is larger for incidental learning (Bertsch et al., 2007), participants were not informed about the memory test but were told that the second session would be similar to the first one. In the testing phase, memory for the items was tested with direct and indirect tests. Both sessions were held in the same dimly lit behavioral testing room and with the same apparatus. Before the start of each session, participants were handed written instructions which they summarized verbally for the experimenter to make sure they had understood everything correctly. While the experiment was running, the experimenter sat in an adjacent room, seeing a display with the identical stimulation provided to the participant, and jotting down their oral responses.

Learning phase. In the learning phase (see Figure 1), participants were informed that there would be two versions of trials: those in which the correct compound word for the triad

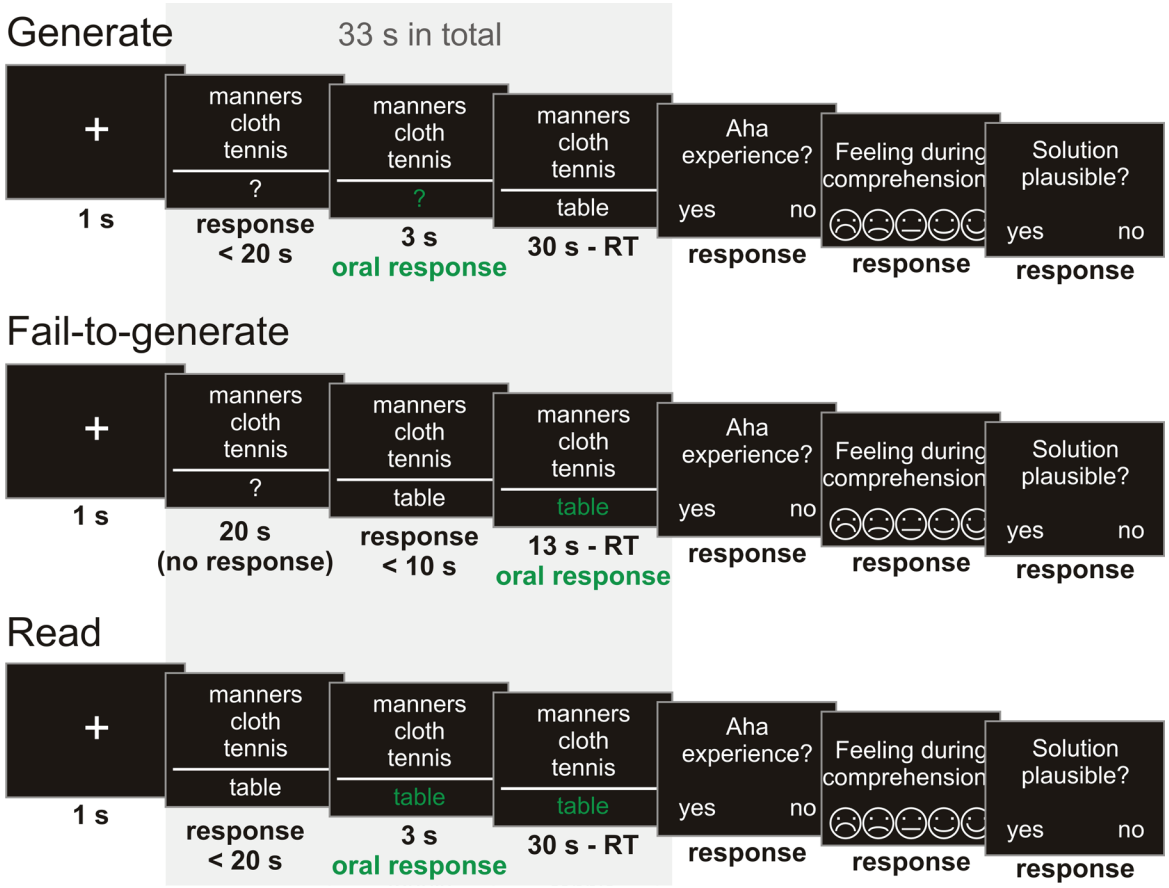

Figure 1.
Learning phase. Example trials of the three generation conditions (generate, fail-to-generate, read). Stimulus presentation had the same duration for all conditions (33 s total) and in each condition, the solution had to be spoken out loud.

had to be found within a time limit of 20 s (chance to generate: generate and fail-to-generate conditions) and others in which the solution word was given and the task was to understand how the solution could be used to create compound words with each of the triad words (read condition). The total presentation duration of the item was held constant (33 s) across the three different levels of GENERATION (generate, fail-to-generate, read) to avoid the duration for processing items as a confound for memory performance differences between conditions.[2]

In trials with a chance to generate the solution, the problem, that is, the three stacked triad words on top of a question mark in place of the solution, was presented for a maximum of 20 s or until a response was made. Participants were instructed to press the spacebar as soon as they came up with the solution word (e.g. "table") and to speak it out loud immediately after button press. The interval for the oral response was 3 s, during which the question mark was highlighted in green. Immediately after, the question mark was replaced by the solution word, which was presented for the remainder of 33 s (which was calculated as 30 s minus the response time as depicted in Figure 1). It was emphasized that they only press the button when they were ready to speak out the answer. Participants were instructed to pay further attention to the item on the screen until item offset.

When no response was made within the time limit (fail-to-generate condition), the solution word was presented in place of the question mark for a maximum of 10 s or until a response was made. Participants were instructed to press the space bar and read the solution word out loud as soon as they understood it. After the button press, the solution word was highlighted in green to indicate the oral response should be made. This display was presented for the remainder of the 33 s, which was calculated as 13 s minus the response time from the first solution display.

In the read condition, the procedure was the same except that there was no search phase, but the solution was presented immediately with the triad. Within the first 20 s, participants indicated their comprehension of the solution by pressing the space bar, followed by reading the solution word out loud. Again, the button press changed the color of the solution word which was highlighted in green. This display remained active for the remainder of the 33 s which was calculated as 30 s minus the response time to the first display of triad and solution.

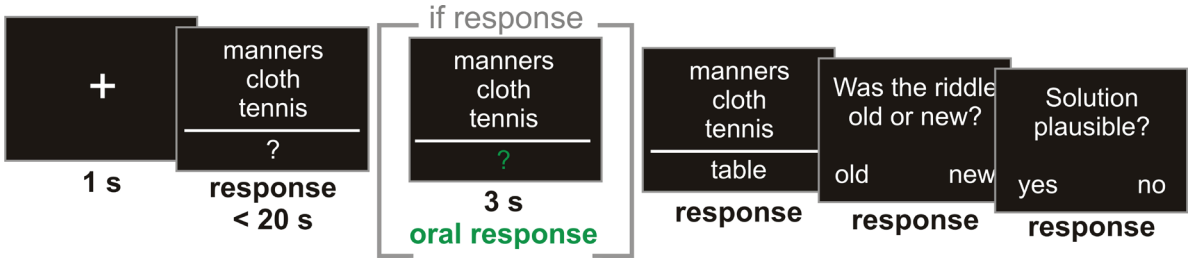

Figure 2.
Memory test of Experiment 1. An indirect (solving old and new items) and a direct memory test (old/new recognition) were applied.

Oral responses had to be made in all three conditions (generate, fail-to-generate, read) to avoid the additional encoding based on speech production as a potential confound for later memory differences. Either the question mark (generate trials) or the solution (fail-to-generate and read trials) changed color from white to green to indicate the interval for speech upon pressing space. Following the item presentation, participants were asked if they had an Aha! experience and answered with "yes" or "no" by pressing the left or right arrow key (assignment of the arrow keys was counterbalanced across participants). After button press, the next display was shown in which participants were asked how they felt when they understood the solution. Participants answered on a 5-point smiley scale as explained in the materials section. They could navigate to the chosen smiley via the arrow keys, with the currently chosen smiley highlighted by a red frame. Pressing the space bar confirmed their choice and led to the next display. Lastly, participants were asked if they thought of the solution as plausible, answering with "yes" or "no" (arrow keys). This button press ended the trial. All trials started with a fixation cross (0.5 s). The task consisted of 135 trials, including three practice trials. Two short breaks of approximately 5 min were made after groups of 45 trials.

Test phase. One week later and approximately at the same time of day as the learning phase, participants were tested for their memory. They were presented with the 135 old CRA items from the learning phase as well as 45 new ones, resulting in a total of 180 trials. The task mainly corresponded to the chance-to-generate condition of the learning phase, that is, the trial timing was similar and participants were instructed to try to solve the problem within a time limit of 20 s (see Figure 2). It was emphasized that participants should not try to recall the solution but rather just try to solve the item regardless of whether they thought of it as old or new (indirect test to measure primarily involuntary memory; Richardson-Klavehn, 2010; Schott et al., 2005). Instead of the Aha! question and the affective rating, participants were then asked if they recognized the item from the last session ("yes" or "no" answer via arrow keys). The button press led to the last display in which participants were again asked if they found the solution plausible ("yes" or "no" answer via arrow keys). There were two short breaks after groups of 60 trials.

Data analysis

All items rated as implausible were discarded from the analysis. First, data were descriptively analyzed in regard to our true independent, manipulated variable, that is, chance-to-solve versus no-chance-to-solve, and the objective dependent variables for memory performance, that is, solution rates of old items and recognition rates of old items. Secondly, inferential statistics were computed based on post-hoc splits of the chance-to-solve condition into generate (correct solutions only, incorrect solutions were discarded from analysis) and fail-to-generate (no solution within the time limit). Additionally, all GENERATION conditions (generate, fail-to-generate, read) were further split into Aha! and no Aha!, based on the subjective ratings of the participants. For this analysis, all incorrect trials were also excluded. Data were analyzed for both memory measures as dependent variables and also for the subjective measure of the emotional response rating as a dependent variable using 3 × 2 repeated-measures ANOVAs with factors GENERATION (generate, fail-to-generate, read) and AHA (Aha!, no Aha!). In case of significant violations of the sphericity assumption, Greenhouse-Geisser (1959) corrected F and p values, the correction parameter epsilon (ε), and uncorrected degrees of freedom are reported. Partial eta squared (η_p^2) and Cohen's d are reported as measures of effect size for all ANOVAs and all t-tests, respectively. For the emotional rating analysis, the median per participant was used before analysis of the means across participants.

RESULTS

Plausibility ratings revealed that .96 (SD = .05) of all old items were rated as plausible in both learning and test phase, and .91 (SD = .07) of all new items were rated as plausible in the test phase. In the learning phase, a solution was generated to .43 (SD = .12) of all chance-to-solve items. Of all generated solutions, .86 (SD = .09) were correct. After discarding all incorrectly generated solutions, .45 (SD = .10) of all chance-to-solve items were correctly generated (from here on, generate refers to correct solutions only) and in the remaining .55 no solution was generated at all (fail-to-generate condition).

Table 2
Mean conditional frequencies of all conditions in Experiment 1 (n = 21).

Condition	Mean	SD
P(generate ∩ no Aha! | chance to generate)	.21	.14
P(generate ∩ Aha! | chance to generate)	.24	.18
P(fail-to-generate ∩ no Aha! | chance to generate)	.14	.16
P(fail-to-generate ∩ Aha! | chance to generate)	.41	.14
P(no Aha! | read)	.69	.35
P(Aha! | read)	.31	.35

The number of items per level of GENERATION was therefore relatively balanced (means of 40.5 items generated, 49.5 items fail-to-generate, 45 items read). The mean frequency of all combinations of GENERATION (generated, fail-to-generate, read) and AHA (Aha!, no Aha!) can be found in Table 2.

Solution rates of old items (indirect memory test measure)

At first, all 21 participants are considered. Old items had a significantly higher solution rate (.64, SD = .09) than new items (.33, SD = .10) for all participants [$t(20)$ = 14.6, p < .001, Cohen's d = 3.197], corroborating that learning occurred. Participants solved .64 (SD = .09) of the chance-to-solve items at test (collapsed over generate and fail-to-generate) and .63 (SD = .10) of the no-chance-to-solve (read) items. However, the difference between the relationships of chance-to-solve and no-chance-to-solve to later memory becomes evident when splitting chance-to-solve for generate and fail-to-generate, as was done in the analysis below.

To investigate the relationship between AHA (Aha!, no Aha!) and GENERATION (generate, fail-to-generate, read) and later solution rates during testing, a 2 × 3 repeated-measures ANOVA was run with the solution rate of old items as dependent variable. For this analysis, the five participants who had zero trials in one of the combinations were discarded. As shown in Figure 3A, there was no effect of AHA, but a main effect of GENERATION [$F(2,30)$ = 26.72, p < .001, η_p^2 = .640]. For the means and standard deviations (SD) that were used in this and all other ANOVAs, please refer to Table 3. Not surprisingly, when participants generated a solution for problems during learning, they had a high probability of solving them again during testing, whereas the solution rate during test was lowest for problems where participants failed to generate a solution during learning. As can be seen in Figure 3A, the read condition scored in between. This pattern was validated via post-hoc tests. Generate (.86, SD = .09) was associated with significantly higher solution rates at test compared to fail-to-generate (.46, SD = .12) [$t(15)$ = 14.9, p < .001, Cohen's d = 3.875] and compared to read (.64, SD = .10) [$t(15)$ = 9.03, p < .000, Cohen's d = 2.278]. Fail-to-generate was associated with significantly lower solution rates at test than read [$t(15)$ = 6.63, p < .001, Cohen's d = 1.674]. It should be noted that even though the lowest solution rate was seen for the fail-to-generate condition it was

Table 3
Mean conditional frequencies of all conditions in Experiment 1 (n = 21). Standard deviations are reported in parentheses.

				Generation	
Experiment	**Test measure**	**Aha!**	**Generate**	**Fail-to-generate**	**Read**
Experiment 1	Indirect	Aha!	0.79 (0.16)	0.48 (0.13)	0.51 (0.31)
(N = 16)		No Aha!	0.82 (0.26)	0.39 (0.26)	0.67 (0.13)
	Direct	Aha!	0.82 (0.20)	0.86 (0.15)	0.74 (0.24)
	Emotional	Aha!	1.09 (0.52)	0.44 (0.63)	0.69 (0.48)
	Response	No Aha!	0.78 (0.55)	0.00 (0.66)	0.31 (0.48)
Experiment 2	Direct	Aha!	0.45 (0.17)	0.35 (0.15)	0.27 (0.32)
(N = 14)		No Aha!	0.38 (0.20)	0.36 (0.21)	0.26 (0.15)
	Emotional	Aha!	1.18 (0.46)	0.46 (0.57)	0.75 (0.55)
	Response	No Aha!	0.68 (0.61)	-0.18 (0.37)	0.07 (0.27)

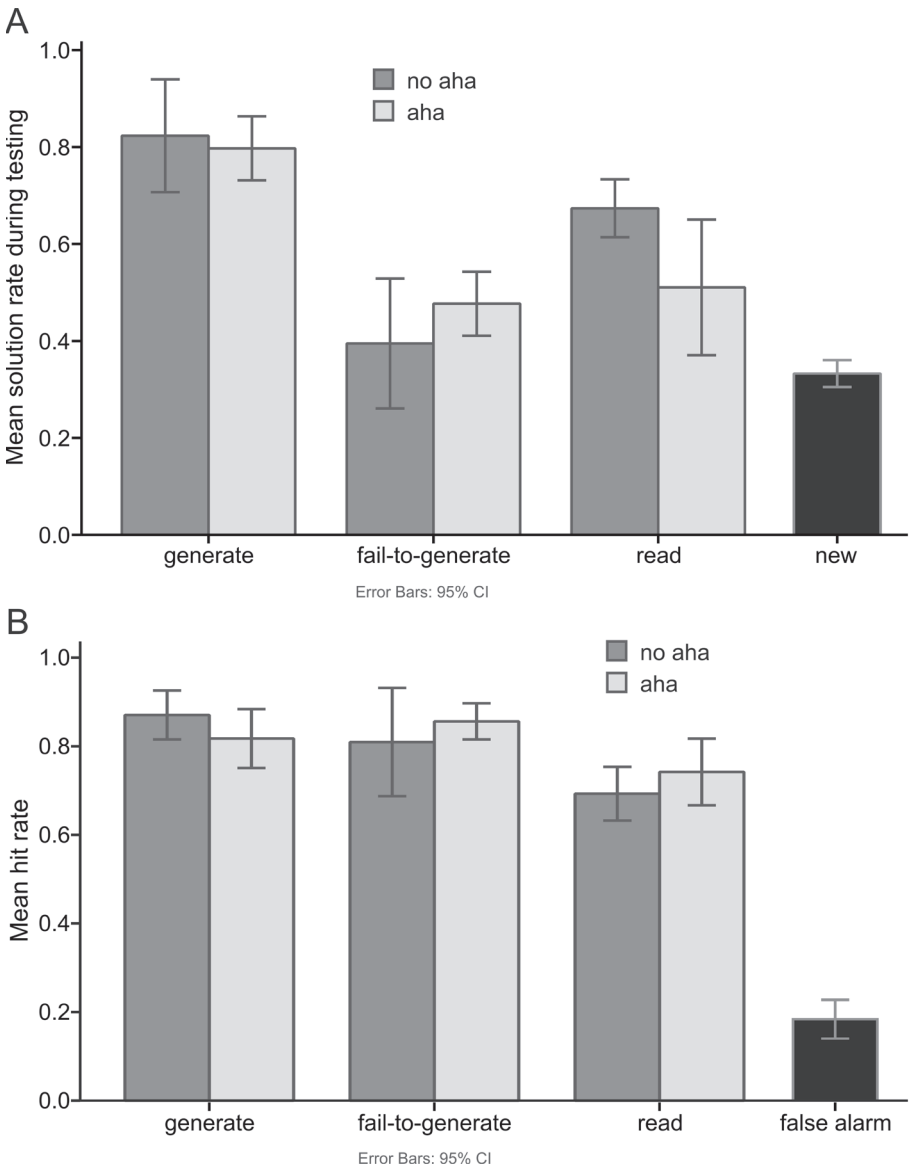

Figure 3.
Memory performance in Experiment 1. A. Mean solution rate of old items during the indirect memory test. The solution rate of new items at test is depicted for comparison. B. Mean hit rate during the direct memory test for solutions presented problems. The false alarm rate (new items incorrectly identified as old) is depicted for comparison.

still marginally higher than the solution rate of new items during testing [$t(15) = 2.02$, $p = .062$, Cohen's $d = 0.657$].

The difference between generate and fail-to-generate suggests a possible selection effect for item difficulty. In other words, it could have been that only easy items and their solutions were learned, and easy items more often landed in the generate condition. This issue was addressed empirically by making a distinction between difficult and easy items based on a median split of the generation rates using our unpublished normative data (i.e., independent data). Indeed, significantly more difficult items landed in the fail-to-generate compared to the generate condition [.66 ($SD = .05$) vs. .29 ($SD = .06$); $t(15) = 19.05$, $p < .001$, Cohen's $d = 6.428$]. However, easy items were not necessarily learned better. To compare the learning rates for difficult and easy items, solution rates at test were compared for old difficult and old easy items, as well as new difficult and new easy items. The learning rate was operationalized as the solution rate for old minus new items for difficult and easy items, respectively. A dependent samples t-test between the learning rate for difficult versus easy items revealed no statistical difference [.14 ($SD = .07$) vs. .17 ($SD = .08$); $t(15) = 1.16$, $p = .263$, Cohen's $d = 0.394$]. Thus, critically, the difference between fail-to-generate and generate cannot simply be attributed to difficult items being learned more poorly.

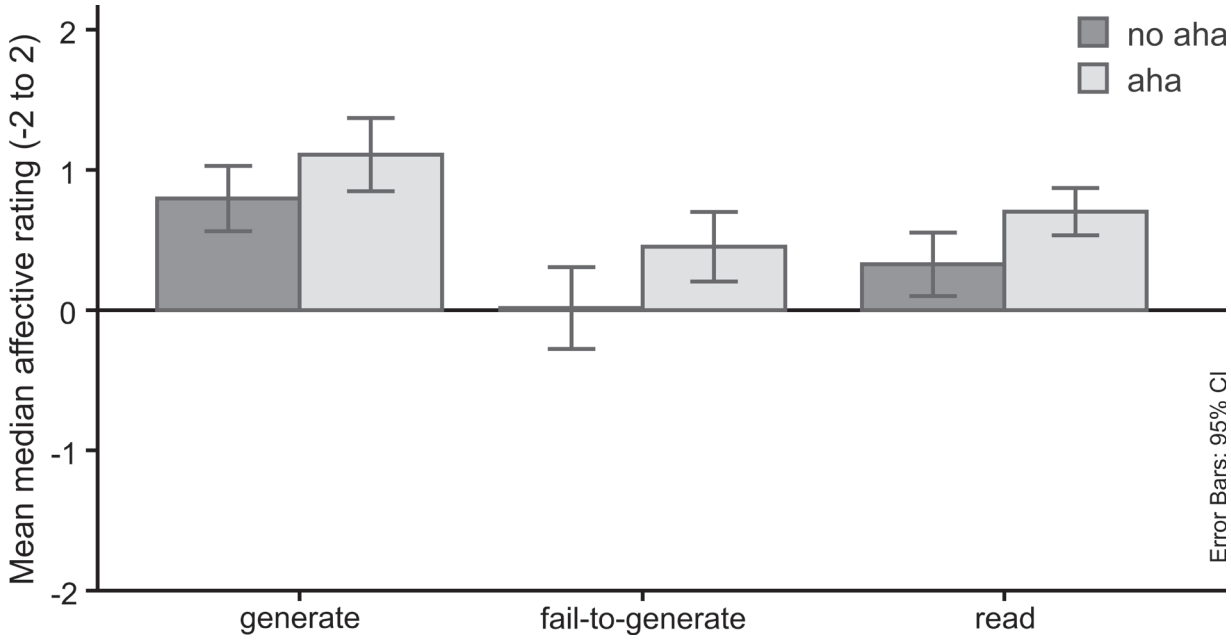

Figure 4.
Emotional response in Experiment 1. Depicted is the mean of the individual median emotional response as measured by a rating on a 5-point scale during the comprehension of the solution.

Recognition memory for solutions presented with problems (direct memory test measure)

In total, .82 (*SD* = .15) of all old items were correctly recognized as such. After subtracting false alarms, discrimination was still at .64 (*SD* = .15). Participants correctly recognized .88 (*SD* = .14) of old chance-to-solve items and .72 (*SD* = .19) of old no-chance-to-solve items. Here, generate and fail-to-generate did not differ in regard to later memory as becomes evident from the analysis below.

As with the solution rate during test, a 2 × 3 repeated-measures ANOVA with factors AHA and GENERATION was run with hit rate as the dependent variable. For this analysis, the five participants with empty cells were again discarded. Again, we found a main effect of GENERATION [$F(2,30) = 5.52$, $p = .009$, $\eta_p^2 = .269$], but no effect of AHA. In contrast to the solution rates of old items, both generate (.85, *SD* = .15) and fail-to-generate (.85, *SD* = .15) led to significantly higher hit rates than read (.69, *SD* = .20) [$t(15) = 4.88$, $p < .001$, Cohen's $d = 1.313$ and $t(15) = 6.74$, $p < .001$, Cohen's $d = 2.050$], but they did not differ from each other [$t(15) = 0.011$, $p = .992$, Cohen's $d = 0.002$] (see Figure 3B). This suggests that attempting to solve a CRAT item enhances later memory for it, even if generation fails, when memory is tested with a recognition memory test. However both the triad and the solution were presented during the recognition memory test, and it is therefore unclear whether only the triad, the solution, or both might have been encoded successfully. Moreover, hit rates appear to be at ceiling for generate and fail-to-generate, which might account for the absence of a statistical difference.

Emotional response. To investigate whether the emotional response differed in quality depending on the generation of the solution or whether an Aha! experience was reported during comprehension, a 2 × 3 repeated-measures ANOVA with factors AHA (Aha!, no Aha!) and GENERATION (generated, fail-to-generate, read) was computed for the median of the affective rating. Again, the five participants with empty cells were excluded from analysis. Main effects of GENERATION [$F(2,30) = 11.60$, $p < .001$, $\eta_p^2 = .436$] and AHA [$F(1,15) = 15.00$, $p = .002$, $\eta_p^2 = .500$] were revealed. As can be seen in Figure 5, Aha! (0.774, *SD* = 0.39) was rated significantly higher than no Aha! (0.51, *SD* = 0.35) [$t(15) = 4.37$, $p = .001$, Cohen's $d = 0.608$]. Generation (0.94, *SD* = 0.47) was also accompanied by a significantly more positive response in comparison to fail-to-generate (0.22, *SD* = 0.54) [$t(15) = 4.11$, $p = .001$, Cohen's $d = 1.029$] and to read (0.50, *SD* = 0.41) [$t(15) = 2.15$, $p = .048$, Cohen's $d = 0.580$], while fail-to-generate received lower affective ratings than read [$t(15) = 2.15$, $p = .048$, Cohen's $d = 0.550$].

DISCUSSION

The results revealed that although generate and Aha! were accompanied by a positive emotional response, only the factor GENERATION showed a significant relationship to later

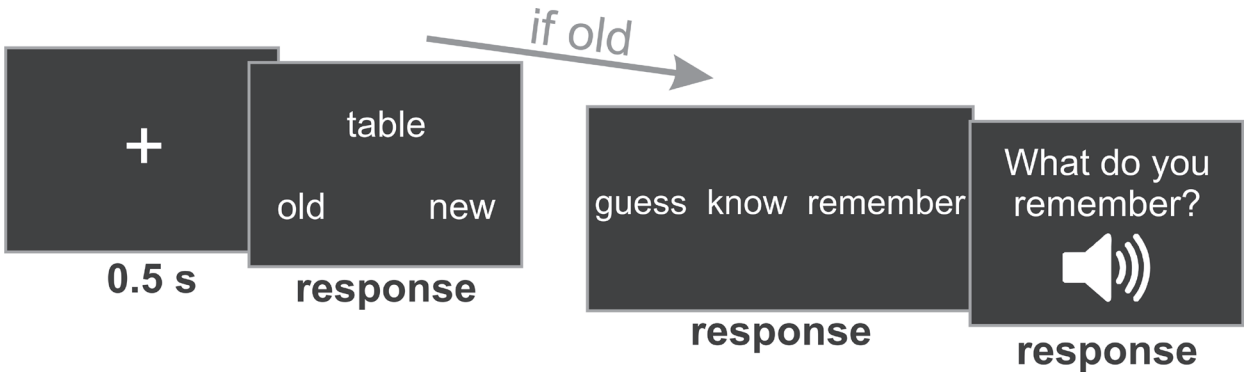

Figure 5.
Memory test of Experiment 2: Exemplary trial.

memory. When an Aha! experience was reported, this was accompanied by a relatively more positive emotional response than no Aha! at all levels of GENERATION. This supports the notion that the Aha! experience induces a positive feeling—even when it cannot be pride of correctly solving a CRAT problem. As for GENERATION, generate was accompanied by the most positive emotional response, followed by read, with the least positive, relatively neutral emotional response for fail-to-generate. While the emotional response to successful generation can probably be explained by pride, the least positive response may be due to disappointment of not being able to solve the item. However, since subjects were not asked about the reasons for these feelings, this result could be investigated in future studies.

Despite the difference between Aha! and no Aha! in regard to the emotional response, AHA did not show a significant relationship to later memory; only GENERATION did. Moreover, depending on whether memory was tested with an indirect or direct test, the ordinal ranking between conditions differed. When the requirement was to solve old items (indirect test), the pattern was generate > read > fail-to-generate, whereas the pattern seen in previous studies was generate = fail-to-generate > read (although in a free recall task; the first current study is apparently the first to use an indirect memory test). When the requirement was to recognize the complete item (problem + solution; direct test), it was generate = fail-to-generate > read, whereas the pattern was generate > fail-to-generate> read in previous studies. On first view, this result suggests that learning from CRAT problem solving is indeed a special case of generation. However, since the apparent equality between generate and fail-to-generate for the direct memory test may have been due to ceiling effects, a second experiment made old/new decisions more difficult by testing the solution word without its triad as context.

EXPERIMENT 2

To eliminate the ceiling effects of the first experiment, the second experiment presented only the solution word as part of a guess/remember/know recognition memory task (see Gardiner & Richardson-Klavehn, 2000). Unfortunately, too few remember responses were given, so remember and know response categories had to be collapsed. The second experiment corresponded to the first one, except for the testing phase (Figure 2). Below only methodological differences to the first experiment are described.

METHODS

Participants

Twenty-two German native speakers (8 male and 14 female) with a mean age of 24.6 years (*SD* = 2.5, range: 21–32) participated in the experiment. Due to empty cells, only 14 participants could be included in the GENERATION × AHA analyses. The remaining participants had a mean age of 24.7 years, and five were male; nine were female.

Task and procedure (testing phase)

Participants were presented with solution words of all old items as well as with solution words of new items[3] and were tested for their recognition memory (see Figure 5). After the presentation of a fixation cross (0.5 s), single solution words were shown in the center of the screen and participants were instructed to decide whether the word was old or new by pressing the left or the right arrow key. The display remained until button press. Participants were instructed to only choose new if they were sure about it and to choose old when they were sure about it as well as when they were insecure. If participants chose new, the next trial started. If they chose old they were asked to indicate whether their decision was just a *guess* or if they *knew* or even *remembered* the solution word. Participants were instructed to choose *know* if they were sure that it was old but did not remember any further context information, and to choose *remember* when they did remember some context information (e.g., their thoughts when they saw the solution in the learning phase or even the related triad). They should

choose *guess* if they could not categorize the item into one of the other categories. Participants made their choice by pressing the right, down, or left arrow key. If they chose guess or know, the next trial started. If they chose remember, they were asked to describe the remembered information to the experimenter, who wrote it down. By pressing the space bar, participants could end the trial and proceed. Again, 135 trials were presented during encoding and 180 during retrieval. Short breaks were made after groups of 45 and 60 trials respectively.

Analysis

In the second experiment, the recognition rate for solution words was the only measure of memory performance. It was planned to analyze the relationship between GENERATION (generated, fail-to-generate, read) × AHA (Aha!, no Aha!) for remember and know responses separately. However, this analysis was not possible because remember responses were, unexpectedly, scarce (possibly due to the long retention interval and the use of single-word test stimuli). Remember and know responses were therefore collapsed, resulting in only one measure of recognition memory for old items (i.e., correct "old" responses). False alarms (i.e., incorrect "old" responses to new items) were treated correspondingly. Guess responses were excluded from the analysis, because they may reflect different decision strategies than true recognition (Gardiner & Richardson-Klavehn, 2000). Also, as in Experiment 1, items rated as implausible and incorrect generations were excluded from analysis.

RESULTS

Plausibility ratings revealed that .93 (*SD* = .05) of all old items were rated as plausible during the learning phase and included in the analysis. In the learning phase, a solution was generated to .36 (*SD* = .06) of all chance-to-solve items. Only .05 (*SD* = .05) of all chance-to-solve items were solved incorrectly. After discarding all incorrect generations, .41 (*SD* = .05) of all chance-to-solve items were correctly generated.

Recognition memory for solutions (direct memory test measure)

Mean hit rate was .59 (*SD* = .23). Corrected for false alarms, discrimination was still at .38 (*SD* = .18), although considerably lower than that in Experiment 1. Of all chance-to-solve

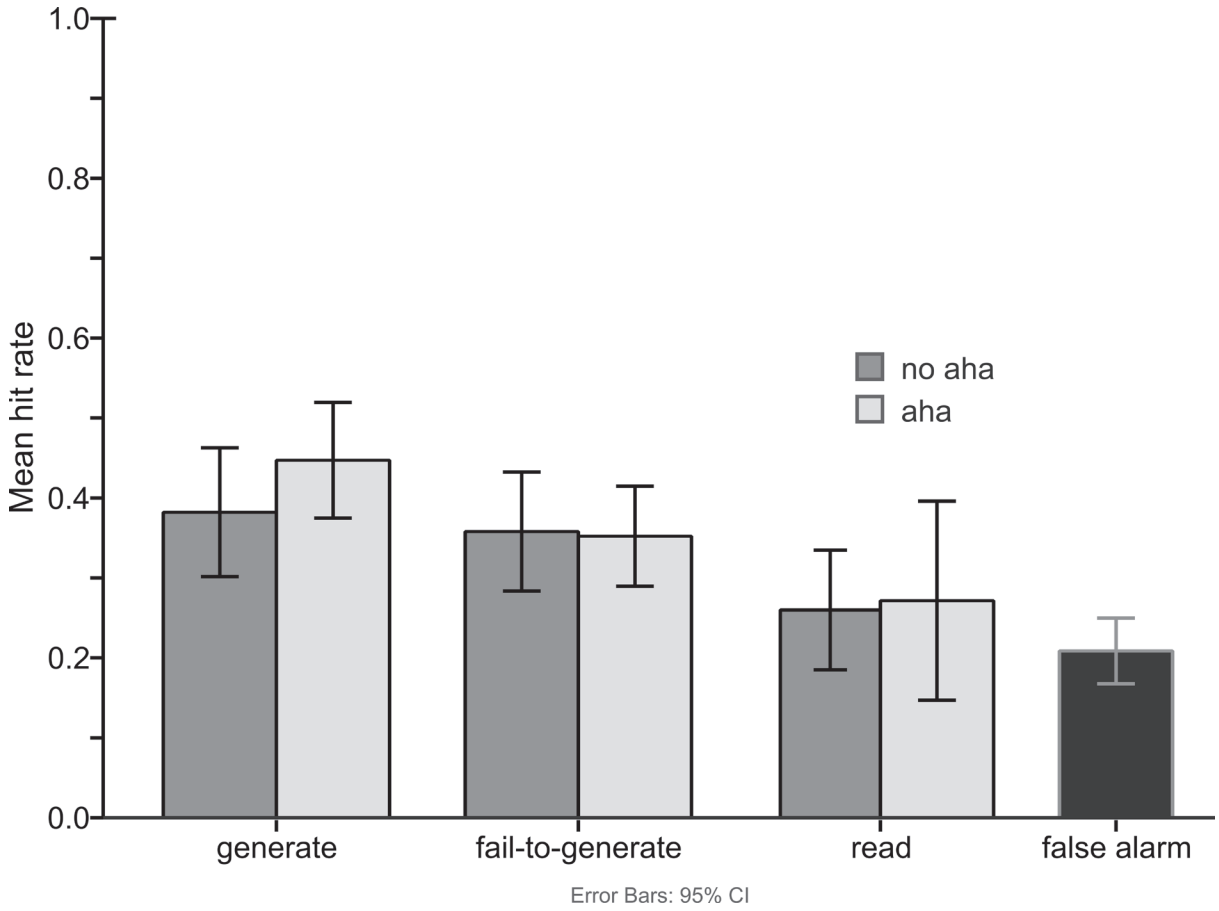

Figure 6.
Memory performance in Experiment 2. Mean hit rate during the direct recognition memory test for solutions. The false alarm rate (new items incorrectly identified as old) is depicted for comparison.

items (generate, fail-to-generate), .66 (*SD* = .22) and .48 (*SD* = .29) of all no-chance-to-solve items (read) were correctly recognized as old. When splitting chance-to-solve into generate and fail-to-generate, it becomes evident that those conditions differ significantly in regard to later recognition memory, as revealed in the analysis below.

A 2 × 3 repeated-measures ANOVA with factors AHA (Aha!, no Aha!) and GENERATION (generate, fail-to-generate, read) was run with hit rate as dependent variable. Partially replicating the result of Experiment 1, only a main effect of GENERATION was found [$F(2,26) = 5.68$, $p = .013$, $\eta_p^2 = .304$, $\varepsilon = .871$]. As can be seen in Figure 6, read (.26, *SD* = .15) differed significantly from generate (.42, *SD* = .15) and fail-to-generate (.34, *SD* = .14) as revealed by post-hoc dependent-samples t-tests [generate vs. read: $t(13) = 3.96$, $p = .002$, Cohen's $d = 1.060$; fail-to-generate vs. read: $t(13) = 2.10$, $p = .056$, Cohen's $d = 0.563$]. In contrast to Experiment 1, there was also a significant difference between generate and fail-to-generate [$t(13) = 2.21$, $p = .046$, Cohen's $d = 0.592$].

Emotional response

The same 2 × 3 repeated-measures ANOVA as above with the median of the affective rating as dependent variable for individual participants was computed to investigate whether the emotional response differs in quality depending on AHA and GENERATION. Only the 14 participants without empty cells were included. As shown in Figure 7, and generally replicating the effect pattern from Experiment 1, there were main effects of GENERATION [$F(2,26) = 17.47$, $p < .001$, $\eta_p^2 = .573$] and AHA [$F(1,13) = 58.60$, $p < .001$, $\eta_p^2 = .818$], with higher ratings for generate (0.93, *SD* = 0.48) than fail-to-generate (0.14, *SD* = 0.38) [$t(13) = 4.58$, $p = .001$, Cohen's $d = 1.644$], higher ratings for fail-to-generate than for read (0.41, *SD* = 0.32) [$t(13) = 2.26$, $p = .042$, Cohen's $d = 0.767$], and higher ratings for Aha! (.80, *SD* = .34) than no Aha! (.40, *SD* = .33) [$t(13) = 5.69$, $p < .001$, Cohen's $d = 1.168$].

DISCUSSION

While emotional response showed the same independent relationships to GENERATION and AHA as in Experiment 1, the direct memory test used in the current experiment revealed different results than the direct memory test in Experiment 1. First, there was no longer a ceiling effect. Second, not only did generate and fail-to-generate differ significantly from read, but also from each other, leading to the pattern: generate > fail-to-generate > read. This is the same pattern that one would expect from Slamecka's and Fevreiski's study on the classical generation effect (Slamecka & Fevreiski, 1983). It further suggests that recognition performance is not dependent on the exposure time to the solution during encoding, which was longest for read (presented for the whole 33 s), shortest for fail-to-generate (presented for 13 s), and medium for generate (presented for approximately 25 s, depending on the solution time, which had a mean of 8 s).

GENERAL DISCUSSION

THE GENERATION EFFECT IN PROBLEM SOLVING

The current study investigated the relationship between problem solving and long-term memory formation in a verbal problem-solving task. The main focus was on the generation

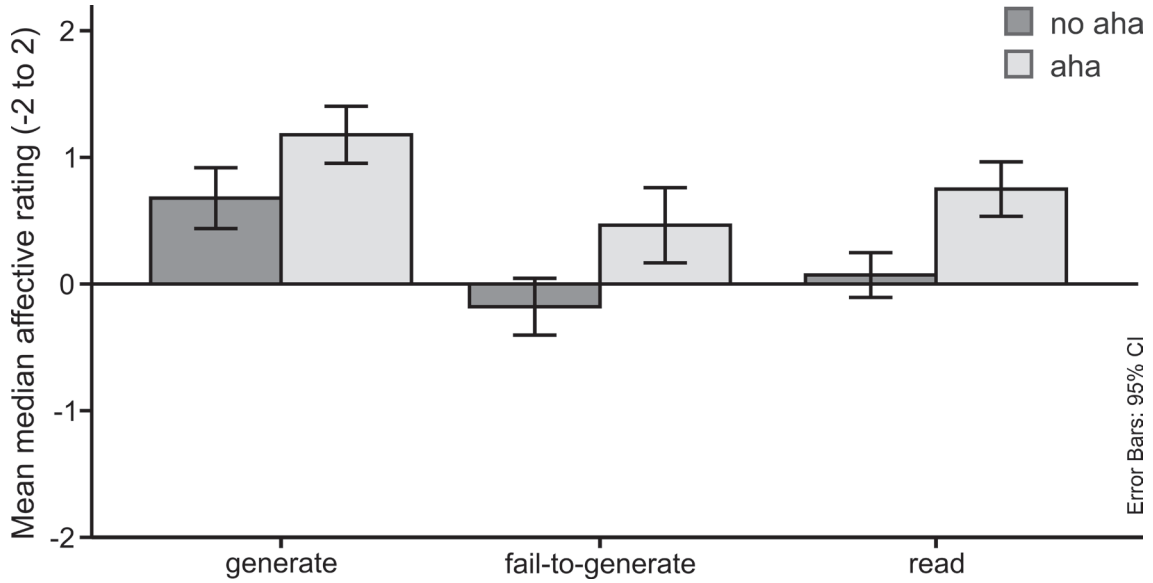

Figure 7.
Emotional response in Experiment 2. As can be seen, the pattern generally corresponds to the one from Experiment 1.

effect, a beneficial effect of generating solutions on later memory, and whether the generation effect in problem-solving tasks that are used to study insight (using the example of the CRAT) represents a special case of the generation effect as observed in traditional memory studies. Memory was compared for generate, fail-to-generate, and read conditions, and in relation to subjective reports of Aha! experiences. In addition, we investigated whether comprehending the solution differed not only in terms of later memory, but also in terms of the emotional response, depending on the generate /fail-to-generate/read and the Aha!/no Aha! variables.

Whereas previous studies showed that both the generation of a solution and the subjective feeling of Aha! were positively related to later memory performance, in this study a relationship was only found for generation. More specifically, generating a solution was associated with the best later memory performance as measured by both indirect and direct memory tests. Memory performance for fail-to-generate and read items differed depending on the type of test. The overall pattern for the indirect test (solution rate of old items) in Experiment 1 was generate > read > fail-to-generate, whereas for the direct memory test (old/new recognition of the solution) it was generate > fail-to-generate > read (when ceiling effects were eliminated in Experiment 2). As summarized in the "Introduction," a generation study in which people were asked to generate the opposites of words compared fail-to-generate with generate and read conditions and reported that fail-to-generate is equal to generate for the free recall of the target item (generate = fail-to-generate > read), but not for recognition memory, where fail-to-generate was associated with a poorer performance (generate > fail-to-generate > read) (Slamecka & Fevreiski, 1983). Slamecka and Fevreiski interpreted this finding as a result from fail-to-generate corresponding to incomplete generation, whereby the item's semantic attributes were activated similarly as in a generate condition. Successful generation is considered to consist of two stages: generating adequate semantic attributes of the solution and arriving at the respective lexical entry that also includes surface features of the word. Since recognition memory relies not only on semantic, but also on surface and processing, fail-to-generate falls short in this regard compared to generate. In regard to CRAT problem solving, although semantic-associative processing in the case of fail-to-generate may not lead closer to the solution, it facilitates later recognition of the solution compared to read, where only a minimum of semantic processing is required to comprehend the presented solution.

In our indirect memory test, that is, trying to solve old items again at test, no memory advantage for fail-to-generate compared to read was found. In fact, items whose solutions had failed to be generated during encoding had a poorer solution performance at test than items whose solutions were read during encoding. These results are consistent with the idea that failing to generate the solution on CRAT problems may lead to interference between the correct solution and the incorrect solutions processed during failed problem-solving attempts (Kane & Anderson, 1978). The memory advantage for solving old CRAT items again when they were previously successfully generated, in comparison to fail-to-generate, is in line with the results of Dominowski and Buyer (2000). However, that study did not include a read condition.

Our findings are in support of the notion of Metcalfe (1986a), who reported considerable differences in the generation process for problem-solving tasks in comparison to classical generation tasks: While classical generation tasks require a search in memory for a specific solution (often a word), problem solving usually requires the construction of a novel solution that is only based on memory content. This may be one reason for differences in the generation effect in problem solving and more classical generation tasks.

THE SUBJECTIVE AHA! EXPERIENCE

Both the subjective Aha! experience and generation were independently associated with relatively more positive emotional responses. As Sternberg (1969) suggested, such additivity might reflect different cognitive processes. Therefore, one interpretation of both generation and Aha! being independently correlated with more positive responses is that different emotions may contribute to the positive feeling. This would also support the notion that the positive emotion which accompanies the Aha! experience is not the pride of solving the problem (Gruber, 2005; Topolinski & Reber, 2010). It has even been reported that this emotional response precedes the finding of a solution (Gruber, 2005). However, whether this positive emotional response truly is something other than pride needs to be investigated in studies that focus more on subjective reports such as those used by Danek and colleagues (2014). Nevertheless, only generation but not the Aha! experience was accompanied by enhanced later memory. The absence of a significant relationship between Aha! and later memory is in contrast with the results of Danek et al. (2013a) and Kizilirmak et al. (2015). A main difference between those two studies and the current one is the stimulus material. Danek et al. used magic tricks and Kizilirmak et al. used Mooney images, that is, their stimulus material was of a more visual nature and may therefore have been more emotionally intriguing. The positive relationship of Aha! to later memory was larger for the presented videos of magic tricks than for the black and white Mooney images (Danek et al., 2013a; Danek, Fraps, von Müller, Grothe, & Öllinger, 2013b; Kizilirmak et al., 2015). It would be plausible that positive relationship between Aha! and the emotional response in the present study, which was also only half the size of the one for Mooney images (Kizilirmak et al., 2015),

was too weak to show a significant relationship with memory performance one week later. However, since the named studies (Danek et al., 2013a; Kizilirmak et al., 2015) may be the only two to have investigated the relationship between the subjective feeling of Aha! and later memory, more studies are needed to investigate the relationship between the type of problem-solving task, the emotional response to the solution, and later memory performance.

LIMITATIONS AND DIFFERENCES FROM PREVIOUS STUDIES

A main limitation of the current study is that the problem-solving task employed here is relatively artificial. It is a task very well suited for laboratory experiments in which many trials of the same type are needed. However, in comparison with insights experienced in real life, the sudden comprehension of the solution to an individual problem may not be nearly as compelling, and the individual trials may not be nearly as distinct. This hypothesis is consistent with the solution performance during the memory test being much poorer than in other studies on learning from insight, which looked at this measure using a small set of classic problems (Ash & Wiley, 2008) or a single problem such as the nine-dot problem (Ash et al., 2012; Dominowski & Buyer, 2000). In the current study, participants were presented with 135 trials of the same problem type. This way, each trial was not nearly as distinctive as the ones in the cited studies. Distinctiveness is known to be an important factor for later memory performance (Eysenck, 1979).

Another limitation of this study is the difference in exposure time of the solution across different levels of generation. Total trial time was held constant, representing a considerable methodological advantage; however, it was not possible to hold both the presentation time of the triad and the presentation time of the solution constant. The differences in exposure time to the solution may, therefore, have influenced later memory. For the direct memory test this potential confound is unproblematic, because the differences in recognition rates were not correlated with the exposure times. At encoding, the exposure-time pattern for the solution was read > generate > fail-to-generate, but on the later memory test the pattern was generate > fail-to-generate > read. For the indirect memory measure, the pattern of solution rates at test was generate > read > fail-to-generate. Although not an identical pattern to the exposure times at encoding, the poor test performance for fail-to-generate might indeed reflect the shortest exposure time at encoding. This possible confounding factor should be evaluated in future studies.

Lastly, although participants were asked to make their Aha!/no Aha! decision for both generated and presented solutions, it may be that there were qualitative differences. Those differences may be confounded with the different levels of generation, and later memory performance. However, the results of the analysis of the emotional response revealed that at least in regard to this qualitative feature, Aha! and generation were independent, suggesting that in this regard the quality of an Aha! experience is not confounded with the level of generation.

CONCLUSIONS

The main finding was that the generation effect for CRAT problems shows a distinctive pattern of memory performance depending on the type of test, that is, solving old items (indirect test) versus recognizing old items or solutions (direct test). While even attempts to generate had an advantage over the read condition in recognition memory performance (generate > fail-to-generate > read), only when generation was successful did it enhance the solution rate of old items during test (generate > read > fail-to-generate). Our results suggest that, contrary to what has been proposed in typical generation studies, fail-to-generate does not necessarily enhance later memory for the solution. Thus, the generation effect in problem-solving tasks such as the CRAT, in which a novel solution has to be constructed based on existing knowledge rather than just retrieving the solution from existing knowledge (typical generation studies), seems indeed to be a special case of the generation effect. Here, being exposed to a solution for a longer time (read) seems to be more helpful for solving the same problem later on than being presented with the solution after a failed attempt at problem solving. Whether this pattern holds also for other problem-solving tasks that have been used to study creativity and insight warrants further investigation. From an educational point of view, this result suggests that different learning strategies should be used depending on how knowledge is tested: When the solution has to be remembered upon the presentation of the problem (especially relevant for a free response test format), trying to solve the problem on one's own during initial exposure to the problem is not always best.

NOTES

1. It should be noted that some classical generation tasks may also be solved by insight. One common task used in insight research and generation effect studies is solving anagrams (Bertsch et al., 2007; Kounios & Beeman, 2009; Metcalfe, 1986b). In generation studies they are simple enough to be always solved, but the way the solution is found may sometimes be through sudden, not gradual, insight.
2. Holding the presentation time constant for both the problem and the solution in the same experiment was not possible. In the case of generation, the solution has to be presented directly after the participant comes up with it, because even if one would continue the display of the problem without the solution, the solution would be active in the participant's mind. Moreover, that procedure could have even more complex consequences, such as the participants having time to question the solution they came up with, which would make any clear conclusions about encoding differences due to the GENERATION conditions themselves impossible.
3. New items consisted of a subsample of CRA items, the same as in Experiment 1. Thus, even though the triad of new items was never presented in Experiment 2, the solution had valid remote associations.

REFERENCES

Anderson, A. K., Yamaguchi, Y., Grabski, W., & Lacka, D. (2006). Emotional memories are not all created equal: Evidence for selective memory enhancement. *Learning & Memory, 13*(6), 711–718. http://dx.doi.org/10.1101/lm.388906

Ash, I. K., Jee, B. D., & Wiley, J. (2012). Investigating insight as sudden learning. *Journal of Problem Solving, 4*(2), 1–27. http://dx.doi.org/10.7771/1932-6246.1123

Ash, I. K., & Wiley, J. (2008). Hindsight bias in insight and mathematical problem solving: Evidence of different reconstruction mechanisms for metacognitive versus situational judgments. *Memory & Cognition, 36*(4), 822–837. http://dx.doi.org/10.3758/MC.36.4.822

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210. http://dx.doi.org/10.3758/BF03193441

Bowden, E. M., & Jung-Beeman, M. (1998). Getting the right idea: Semantic activation in the right hemisphere may help solve insight problems. *Psychological Science, 9*(6), 435–440. http://dx.doi.org/10.1111/1467-9280.00082

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

Bowers, K. S., Regehr, G., Balthazard, C., & Parker, K. (1990). Intuition in the context of discovery. *Cognitive Psychology, 22*(1), 72–110. http://dx.doi.org/10.1016/0010-0285(90)90004-N

Burns, D. J. (1992). The consequences of generation. *Journal of Memory and Language, 31*(5), 615–633. http://dx.doi.org/10.1016/0749-596X(92)90031-R

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428. http://dx.doi.org/10.1037//0033-295X.82.6.407

Cranford, E. A., & Moss, J. (2010). Investigating insight using compound remote associate problems. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1768–1773). Austin, TX: Cognitive Science Society.

Cranford, E. A., & Moss, J. (2012). Is insight always the same? A protocol analysis of insight in compound remote associate problems. *Journal of Problem Solving, 4*(2), 128–153. http://dx.doi.org/10.7771/1932-6246.1129

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

Danek, A. H., Fraps, T., von Müller, A., Grothe, B., & Öllinger, M. (2013b). Working wonders? Investigating insight with magic tricks. *Cognition, 130*(2), 174–185. http://dx.doi.org/10.1016/j.cognition.2013.11.003

Danek, A. H., Fraps, T., von Müller, A., Grothe, B., & Öllinger, M. (2014). It's a kind of magic—What self-reports can reveal about the phenomenology of insight problem solving. *Frontiers in Psychology, 5*(1408), 1–11. http://dx.doi.org/10.3389/fpsyg.2014.01408

Dominowski, R. L., & Buyer, L. S. (2000). Retention of problem solutions: The re-solution effect. *American Journal of Psychology, 113*(2), 249. http://dx.doi.org/10.2307/1423730

Eysenck, M. W. (1979). Depth, elaboration, and distinctiveness. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 89–118). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gardiner, J. M., Java, R. I., & Richardson-Klavehn, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 50*(1), 114–122. http://dx.doi.org/10.1037/1196-1961.50.1.114

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness and Cognition, 7*(1), 1–26. http://dx.doi.org/10.1006/ccog.1997.0321

Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 229–244). New York: Oxford University Press.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95–112. http://dx.doi.org/10.1007/BF02289823

Gruber, H. E. (2005). Insight and affect in the history of science. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 397–431). Cambridge: MIT Press.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57*(6), 787–795. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7651803

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., . . . Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology, 2*(4), E97. http://dx.doi.org/10.1371/journal.pbio.0020097

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology, 70*(4), 626–635. http://dx.doi.org/10.1037//0022-0663.70.4.626

Kizilirmak, J. M., Galvao Gomes da Silva, J., Imamoglu, F., & Richardson-Klavehn, A. (2015). Generation and the subjective feeling of "aha!" are independently related to learning from insight. *Psychological Research.* http://dx.doi.org/10.1007/s00426-015-0697-2

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6), 1534–1555. http://dx.doi.org/10.1037/0278-7393.25.6.1534

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

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

McElroy, L. A., & Slamecka, N. J. (1982). Memorial consequences of generating nonwords: Implications for semantic-memory interpretations of the generation effect. *Journal of Verbal Learning and Verbal Behavior, 21*(3), 249–259. http://dx.doi.org/10.1016/S0022-5371(82)90593-X

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220–232. http://dx.doi.org/10.1037/h0048850

Metcalfe, J. (1986a). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(2), 288–294. http://dx.doi.org/10.1037/0278-7393.12.2.288

Metcalfe, J. (1986b). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(4), 623–634. http://dx.doi.org/10.1037/0278-7393.12.4.623

Richardson-Klavehn, A. (2010). Priming, automatic recollection, and control of retrieval: Toward an integrative retrieval architecture. In J. H. Mace (Ed.), *The act of remembering: Toward an understanding of how we recall the past* (pp. 111–179). Oxford, UK: Wiley-Blackwell.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology, 39*(1), 475–543. http://dx.doi.org/10.1146/annurev.ps.39.020188.002355

Schott, B. H., Henson, R. N., Richardson-Klavehn, A., Becker, C., Thoma, V., Heinze, H.-J., & Düzel, E. (2005). Redefining implicit and explicit memory: The functional neuroanatomy of priming, remembering, and control of retrieval. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1257–1262. http://dx.doi.org/10.1073/pnas.0409070102

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior, 22*(2), 153–163. http://dx.doi.org/10.1016/S0022-5371(83)90112-3

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory, 4*(6), 592–604. http://dx.doi.org/10.1037//0278-7393.4.6.592

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30*, 276–315. http://dx.doi.org/10.1016/0001-6918(69)90055-9

Topolinski, S., & Reber, R. (2010). Gaining insight into the "aha" experience. *Current Directions in Psychological Science, 19*(6), 402–405. http://dx.doi.org/10.1177/0963721410388803

Yonelinas, A. P., & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: An emotional binding account. *Trends in Cognitive Sciences, 19*(5), 259–267. http://dx.doi.org/10.1016/j.tics.2015.02.009