THINKING & REASONING
2020, VOL. 26, NO. 4, 552–580
https://doi.org/10.1080/13546783.2019.1685001

Routledge
Taylor & Francis Group

Why do people fail to see simple solutions? Using think-aloud protocols to uncover the mechanism behind the Einstellung (mental set) effect

Christine Blech[a], Robert Gaschler[a] and Merim Bilalić[b]

[a]Department of Psychology, FernUniversität, Hagen, Germany; [b]Department of Psychology, Northumbria University Newcastle, Newcastle upon Tyne, UK

ABSTRACT

Einstellung (mental set) effects designate the phenomenon where established routines can prevent people from finding other, possibly more efficient solutions. Here we investigate the mechanism behind this phenomenon by using Luchins' classical water jug paradigm with concurrent verbalization. We find no difference in the extent of the Einstellung effect between the group which was instructed to think aloud during the problem solving and the group which was thinking silently. The think-aloud protocols indicate that the participants who exhibited the Einstellung effect repeatedly attempted to solve the water jug problem by using variations of the previously successful method which had been rendered inappropriate in the final problem. Our study underlines the usefulness of the think-aloud technique in tracking the cognitive processes. More importantly, it demonstrates how, once thought has been activated, it may bias subsequent dealings with new situations, even in the face of repeated failure that people experience in the Einstellung situations.

ARTICLE HISTORY Received 3 July 2018; Accepted 20 October 2019

KEYWORDS Einstellung (mental set) effect; fixation; water jug task; thinking aloud; cognitive processes

Have you ever wondered why many people tend to stick to routines or habits? One reason may be that learning new practices implies some kind of investment or effort. Consider, for example, the changeover from text processing via typewriter to text processing by means of personal computers (Ceruzzi, 2003). Using this new technique certainly presented a challenge to office workers of the 1980s and 1990s. The investment in training sessions and equipment purchase had to be weighed against long-term

CONTACT Christine Blech christine.blech@fernuni-hagen.de Department of Psychology, FernUniversität in Hagen, Universitätsstraße 33, D-58097 Hagen, Germany

benefits such as more efficient writing and data exchange. In other instances, a person may simply not be aware of a new and promising option. As Luchins (1942) puts it, if someone prefers a well-known routine although a simpler, more economic approach to a task or problem would be applicable, there must be 'blindness' towards alternative solutions. In line with Luchins, we will apply the term *Einstellung* (mental set) effect to this phenomenon. While many routines are reasonable aids to budget our mental, physical, and general resources, in other situations unreflective use of routines may not only prevent us from finding more suitable behaviors. It may even be detrimental, such as when wrongly diagnosing a rare, severe disease as a common, harmless indisposition (Crosskerry, 2003). Here we investigate the reasons for the inability to spot alternatives once the mind has adopted a preferred way of dealing with the situation.

The Einstellung effect and water jug paradigm

Luchins' (1942) original paradigm to demonstrate Einstellung was the water jug task, in which participants are to combine the contents of virtual jars to attain a fixed target amount. Figure 1 illustrates the task rules by means of a sample problem: The target amount of 100 units (say, liters), has to be produced by a linear combination of the three jugs A, B, and C. Hence the values of 21, 127, and 3 can be added and subtracted from one another so that the equation exactly matches the value of 100 (Figure 1A). In this example, the correct solution is: B – A – 2C, that is, filling the largest jug (127) full, then filling the amount of the medium jug (21) from the largest jug (127 – 21), and finally using the rest in the largest jug (106) to fill the smallest jug twice ($2 \times 3 = 6$, $106 - 6 = 100$). Luchins' participants were confronted with five introductory problems, which we call here Einstellung problems, of the same solution scheme: B – A – 2C. Two so-called critical problems (Figure 1B) followed, which we call here 2-solution problems because two solutions are possible: the old Einstellung, or E-solution, B – A – 2C, as well as the easier direct or D-solution A – C. Finally, participants were presented with the extinction problem, which we call here a 1-solution problem because only the D-solution, A – C, was possible (Figure 1C).

Luchins found that most people do not see the shorter solution in the 2-solution problems and instead go along with the previously successfully applied E-solution. More surprisingly, most of the participants pronounced the extinction 1-solution problem, where the E-solution was not possible, to be unsolvable. Luchins' remarkable E-effect demonstration has been replicated many times directly (for reviews, see Luchins & Luchins, 1994; Schultz & Searleman, 2002) but also in a variety of other formats (e.g., Chen

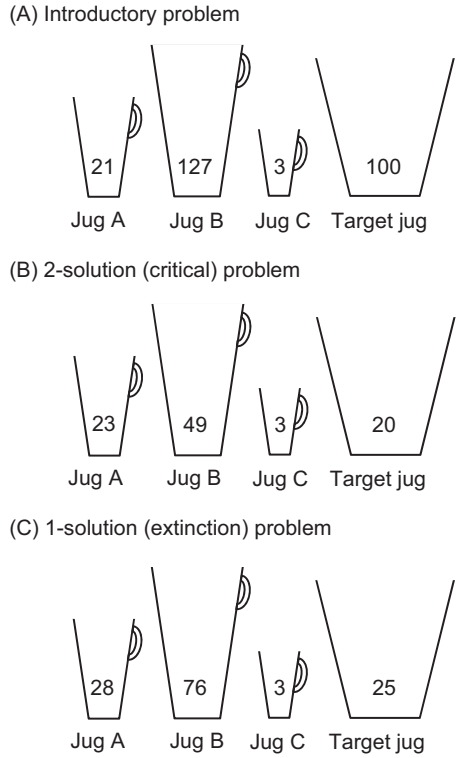

Figure 1. Illustration of the water jug paradigm (Luchins, 1942) in its implementation on PowerPoint slides. Panel (A) depicts the first of five routine or Einstellung problems conforming to the solution scheme B – A – 2C. Panel (B) depicts the first of two critical (2-solution) problems which can be solved either by the Einstellung scheme B – A – 2C or by the short, direct solution A – C. Panel (C) depicts the extinction (1-solution) problem allowing only for the direct solution A – C.

& Mo, 2004; Delaney, Ericsson, & Knowles, 2004; Lovett & Anderson, 1996; Woltz, Gardner, & Bell, 2000).

Mechanisms behind the Einstellung effect: the impact of memory activation

Despite the multitude of research on the water jugs paradigm, it is unclear why the Einstellung effect occurs. Previous explanations (Atwood & Polson, 1976; Delaney et al., 2004; Greeno, Magone, & Chaiklin, 1979) use the classical system production framework (Anderson, 2013; Newell & Simon, 1972), which assumes that participants develop procedures for dealing with certain situations. Procedures, collections of rules to be applied in certain situations, in this particular context would be the common way of dealing with the water jug problem (e.g., the B – A – 2C method). As the procedure

becomes stronger with more frequent use in the introductory problems, the participants become increasingly fixated on that particular way of dealing with the problem (for more details, see Bilalić, McLeod, & Gobet, 2008a).

This explanation may indeed account for people's use of the old E-method in the two critical (2-solution) problems. The old procedure is readily available and, most importantly, it still works perfectly, thus giving no feedback that something is amiss. However, the procedural explanation is hardly applicable to the final extinction (1-solution) problem. The old procedure (E-method) is not available in the final problem. The new solution is not only the shorter and cognitively less demanding one, but it is the only way to solve the problem. Therefore, a conscious or unconscious aversion to investing effort and applying the well-learned procedure is unlikely to be the only explanation for participants' inability to break through the mental set.

Beyond memory activation: the interplay between memory, attention, and perception

A possible clue about the mechanism behind the inability to find a simple solution in the water jug paradigm is provided by a study using a similar paradigm with a different population of participants. Our previous study (Bilalić, McLeod, & Gobet, 2008b) measured the eye movements of expert chess players while they solved a chess problem that could be solved using a well-known solution that would immediately come to players' minds. However, similar to the water jug paradigm, the chess problem could also be solved using a shorter, more optimal solution (a 2-solution problem). The experts immediately spotted the familiar, suboptimal solution, but regularly failed to find the shorter solution despite claiming to be looking for one that was better. Eye movement demonstrated that experts mostly fixated the features related to the old familiar solution even when they claimed that they were looking for new ones. Therefore, the alternatives the experts had investigated were closely related to the initial, familiar solution and consequently prevented them from finding the better solution. In other words, their new solutions were variations on the old method.

The same Einstellung mechanism has recently been replicated in similar situations in the same domain (chess – Sheridan & Reingold, 2013) but also extended to another verbal domain with laypeople (anagrams – Ellis & Reingold, 2014). The discovered mechanism may also be at work in the water jug paradigm. The first idea that comes to mind, the previously acquired and vastly successful E-solution, directs attention toward the aspects of the problem which are related to the E-method. The E-method

in the water jug paradigm is directly related to the use of Jug B (see Figure 1C). Jug B is the starting point of the E-method, which is natural given that Jug B has the biggest capacity in all problems. Most problems cannot be solved by starting with other jugs. The participants will then start to solve the final 1-solution (extinction) problem using the E-method as well, that is, starting with Jug B. As the first attempt inevitably fails, since it is impossible to solve the final problem when starting with Jug B[1], the initial idea (E-method) will still guide their subsequent search. The participants will continue to start with Jug B in their new attempts, trying to refine the method that had worked so well in the past. In other words, participants' new attempts will inevitably represent a variation on the E-solution theme. This will further reinforce their mental set and make real alternatives (e.g., those that start with a jug other than jug B) even more remote.

The proposed Einstellung mechanism bears similarities to a number of theories that assume that memory activation is behind fixation phenomena such as Einstellung effect (e.g., Crilly & Cardoso, 2017, Nickerson, 1998; Nijstad & Stroebe, 2006; Smith, 1995a). They all explain why the old solution is the first one that comes to mind and, consequently, why the initial fixation occurs. However, the Einstellung explanation (see also, Bilalić, McLeod, & Gobet, 2010; Bilalić & McLeod, 2014) goes further because it postulates how the subsequent search for new solutions is influenced by the initial idea. The memory activation is just the beginning of a pernicious cycle where subsequent attention leads to further perceptions of the elements related to the initial idea. This in turn feeds back the initial (inappropriate) memory activation. Although people believe that they are looking for new solutions, they will inevitably be trying to refine the initial way of dealing with the situation.

Think-aloud protocols for tracing mental representations in E-Effect

Here we wanted to check whether the same mechanism uncovered in experts is the driving force behind the original Einstellung phenomenon. Instead of employing the eye tracking technique used in previous studies (Bilalić et al., 2008b; Ellis & Reingold, 2014; Sheridan & Reingold, 2013), here we use think-aloud protocols (Ericsson & Simon, 1993). Thinking aloud involves participants verbalizing their thoughts while working on a specific task (Ericsson, 2006). Think-aloud protocols have a long tradition in problem solving and are one of the main process tracking techniques (Ericsson & Simon, 1993).

[1]It is possible to solve the extinction (1-solution) problem using a rather cumbersome method of filling Jug B full, and filling Jug C 17 times ($76 - 17 \times 3$).

In the context of the set effect, the verbal protocols can provide information about the involvement of the irrelevant jug B in the final extinction (1-solution) problem. This information can tell us whether an attempt starts with Jug B, when one attempt starts and when it ends, how many attempts were tried and what exactly was their content. This would in turn enable us to draw conclusions about the proposed underlying mechanism. It may also prove superior to eye tracking in this particular context because we can differentiate between solution attempts from verbal protocols. This is not easily done with eye tracking. For examples, a long continued fixation period does not reveal whether one or several attempts are undertaken. Similarly, a (longer) switch between visual areas of interest can possibly, but does not necessarily, mark the transition to a new solution attempt.

On the other hand, verbal protocols might potentially be reactive, affecting the E-effect: As a person is compelled to stick to the verbal and analytical level of thinking, the thinking-aloud procedure might hamper the very mechanisms which are required in order to offset the Einstellung. If the set effect arises from adverse activation, pausing (Penney, Godsell, Scott, & Balsom, 2004) or distracting activities would be appropriate to stop thoughts concerning the problem (Sio & Ormerod, 2009). Verbalization might shield the person from such activities that otherwise might stop the mental set.

In the literature, there is evidence both in favor of and against the reactivity of think-aloud methods. The research on insight problem solving provides evidence that thinking aloud may actually hamper performance. Schooler, Ohlsson, and Brooks (1993) demonstrated that verbalization of thought prevents people from finding insight solutions, unlike in non-insight problems where the verbalization has no adverse effects. The assumption underlying this claim of *verbal overshadowing* is that insight problem solving and non-insight problem solving differ in their very nature In terms of the *special-process theory* (cf. Seifert, Meyer, Davidson, Patalano, & Yaniv, 1995), solving an insight problem involves mechanisms of cognitive restructuring, which are beyond the scope of verbal reporting. Hence, any form of addressing the problem in a verbal manner implies irrelevant mental action, barring the way to the solution.

For example, Ball, Marsh, Litchfield, Cook, and Booth (2015) used a set of visual insight problems to demonstrate that overt verbalization in thinking aloud as well as the problem-related inner language, which one would assume to take place in silent control participants (Baddeley, 2007; Baddeley, Gathercole, & Papagno, 1998), lowered solution accuracy and increased solution times. The opposite was found for experimental conditions in which the working memory was supposedly cleared of problem-related inner language by either articulatory suppression or listening to

irrelevant speech (Ball et al., 2015). Acknowledging that the E-effect bears similarities to the insight phenomenon (Öllinger, Jones, & Knoblich, 2008), such as requiring an insight-like breakout from initial (and unhelpful) cognitive schemes, one might expect that think-aloud protocols may increase the E-effect.

On the other hand, an earlier study by Ball and Stevens (2009) with verbal insight tasks, specifically compound remote associates (word tasks involving a shared association among three words), revealed that subjects thinking aloud performed better than subjects under the condition of articulatory suppression and – if the tasks were high in complexity – even better than silent thinkers. Contradictory to the predictions of the special-process theory, the results were in line with the *business-as-usual theory* (e.g., MacGregor, Ormerod, & Chronicle, 2001). Business-as-usual theory assumes that dealing with insight and non-insight problems relies on the same basic cognitive mechanisms, such as, e.g., means-end analysis. This is why thinking aloud can potentially improve insight problem solving by serving as a structuring support (cf. Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Neuman & Schwarz, 1998; Redifer, Therriault, Lee, & Schroeder, 2016).

The reviewed studies also suggest moderating factors which might affect when (overt) verbalization is helpful to a solution process and when it is not. For example, the duration of a task can be crucial, in that thinking aloud impairs, i.e., delays, the early solution process whereas during later stages of problem solving think-aloud subjects and silent thinkers perform almost equally well (Ball et al., 2015; see also the meta-analysis by Fox, Ericsson, and Best (2011) with the general finding that expressing thoughts aloud entailed higher time demands, probably due to the time-consuming activity of speaking). Since one can hardly foresee the exact moment at which potential impairments due to thinking aloud will vanish, in the present experiment on water jug problems we did not only assess the solution accuracy, i.e., whether a person managed to find the proper solution within a fixed and somewhat arbitrary time frame, but also the solution time, in order to allow for process analyses.

Current study

Given the above considerations, it seems clear that we cannot simply assume that think-aloud protocols will illuminate the mechanism behind the Einstellung effect in the water jug paradigm. We have therefore directly included in our study participants who solve the problems while thinking aloud and those who solve the problems silently. Should the thinking mode affect the E-effect, we would expect significant differences between

the two groups in accuracy and time needed to find the D-method in the final 1-solution problem.

Altogether, we have three conditions: (1) the condition where think-aloud and silent participants solve Luchins' classical water jug paradigm, (2) the condition where a new set of participants (also featuring two groups, think-aloud and silent) only solve the introductory Einstellung problems without the critical 2-solution problems, and (3) the control group of participants who are shown the final extinction (1-solution) problem immediately (again two groups, think-aloud and silent). We call the first group E_7 because there are seven Einstellung problems that the participants need to solve before they encounter the final 1-solution problem. Similarly, the second condition is called E_5 as there are only five Einstellung problems. The third condition is called E_0 as the control participants do not need to solve any Einstellung problems. To sum up, participants in all three conditions complete an identical version of the extinction (1-solution) problem. What differs is the number and types of problems that will be presented prior to the extinction problem.

The condition with critical 2-solution problems (E_7) provides participants with more introductory problems and therefore makes them more prone to the E-effect (Luchins, 1942). However, some of the participants will inevitably find the shorter solution in the two critical problems. This would break the established mental set and these participants would then have to be excluded from the analysis in order not to be mixed with the participants who still have an intact mental set. The E_5 condition provides fewer introductory problems, but enables us to use a more efficient experimental design where all participants can be analyzed on the final 1-solution problem.

We assumed that the mental set effect would occur in both E-groups, manifesting itself in more and quicker solutions of the extinction (1-solution) problem in the control E_0 group than in the two Einstellung conditions (E_5 and E_7). Most importantly, we assume that most participants in the E-groups will be using a variation of the old E-method when solving the final 1-solution problem. In other words, their subsequent attempts will still feature Jug B. Hypotheses and research questions were pre-registered online on the Open Science Framework Platform (Blech, Gaschler, & Bilalić, 2018).

Method

Participants

One hundred and eighty-two participants took part in the study. Thirteen participants were excluded from later analyses because they found the Einstellung solution in fewer than three out of the five introductory

problems. Another participant was taken out of the final sample due to a small amount of verbalization in the thinking aloud condition. The final sample consisted of 168 German-speaking participants (85 male, 2 with sex unspecified). The mean age was $M = 35.90$ years ($SD = 12.34$ years) with a range from 14 to 67 years.

There were 28 participants in E_5-silent condition, 24 participants in E_5-think aloud, 30 in E_7-silent, 26 in E_7-think aloud, 30 in E_0-silent, and 30 participants in E_0-think aloud condition.

Procedure and material

The participants were recruited by the students from two project courses enrolled in the B.Sc. curriculum at the FernUniversität in Hagen. Each of the thirty students recruited and tested six participants (one person for each cell of the two-by-three design) from among their personal acquaintances. The participants were allocated randomly to the six conditions. The recruited persons were blind as to the background of the study.

The test sessions took place individually in quiet rooms, shielded from everyday disturbances such as telephone calls. Participants were given a brief overview concerning the topic of the study and its approximate duration. A full informed consent form signed by subject and experimenter ensured confidential treatment, anonymized analysis of the data, and the option for participants to withdraw their agreement up to four days after the end of the sampling period.

The main experimental task was presented on a computer screen using PowerPoint slides while the software Audacity recorded the concurrent verbalizations in the think-aloud groups and the verbally expressed solutions in the silent groups via microphone. For the E_7 condition we applied a variant of the full Luchins paradigm, designing five water jug tasks with the three jugs A, B, and C following the B − A − 2C solution (see Figure 1A). These represented the *Einstellung problems*. The sixth and the seventh problem were *critical (2-solution) problems* (see Figure 1B). They were ambiguous, allowing both for the complex Einstellung solution (B − A − 2C) and for a simple direct solution (problem 6: A − C, problem 7: A + C). A final eighth problem, the *extinction (1-solution) problem*, also employed three jugs A, B, and C, and appeared identical in the surface structure, yet the Einstellung solution B − A − 2C was not applicable. Instead the required solution was A − C (Figure 1C). A complete list of all problems used is given in Table A1 in the appendix. The E_5 condition was given the five Einstellung problems and the final extinction (1-solution) problem, but not the critical (2-solution) problems. The control condition E_0 had to answer nothing but the extinction (1-solution) problem. Each problem was shown for a maximum

duration of 180 seconds. After that interval, the PowerPoint slide turned over automatically to the next problem while for faster solutions a self-paced progress was possible. Transitions from one slide to the next were signaled by a cymbal-like tone in order to insert time markers into the later transcripts.

The water jug introduction did not specify the number of jugs to be used, but it made clear that subjects were allowed to use water jugs more than once and that the content had to be transferred completely (without any residue) from one jug into another. For the thinking aloud groups the technique was briefly explained as speaking out loud everything that comes to the subject's mind, and it was practiced with the help of two simplified introductory sample tasks involving merely two jugs rather than three. The silent subgroups were presented with the same tasks. For them, the only verbal expressions allowed and required during the experiment were their statements of the solutions.

After the water jug tasks, participants indicated their sex, age, and educational level in a concluding pencil and paper questionnaire before they were fully debriefed, receiving relevant background information as to the research hypotheses.

Transcription, coding, and dependent variables

For participants from the silent thinking conditions we transcribed the verbally expressed solutions into text documents; for the think-aloud condition we took down the complete progress of verbal utterances, both the final solution as well as intermediate steps. Standardized, common transcription rules were applied (Dresing, Pehl, & Schmieder, 2015). Dialects and idioms were written down as in High German. Time stamps were inserted by number signs ('#') before and after every problem. The solution time for the extinction (1-solution) problem was calculated from the respective time stamps in the protocols. Based on the written transcripts, Einstellung problem responses were coded as correct Einstellung solutions (1) or missing solutions (0). Critical (2-solution) problem responses were classified as Einstellung solutions (1), direct solutions (2) or incorrect or missing solutions (0), and in the extinction (1-solution) problem we distinguished between the correct direct solution (1) vs. no direct solution (0).

Concerning the extinction (1-solution) problems of the think-aloud participants, we conducted a closer analysis of the intermediate solution steps. An independent rater – blind to the background of the experiment – coded the 80 transcripts by segmenting the verbalizations of the extinction (1-solution) problem into one or more solution attempts, depending on the length and the structure. Longer breaks, explicit statements like 'okay, once

again', or restarts that were obvious from the logic of the numerical values indicated separate, successive solution attempts within the extinction (1-solution) task. Each attempt was formalized as an algebraic term in which numerical values corresponding to jug contents were written as letters, e.g., the phrase '[jug] B 76, [...] 28 times 2 equals 56. 20 would remain [...]' was coded as 'B – 2A' since the passage about the remainder of 20 units implies a subtraction. In order to objectify the procedure, a second rater coded a random sample of 20% of the critical (2-solution) problem transcripts (as suggested by Neuendorf, 2002). The interrater reliability was determined by Cohens kappa, yielding $\kappa = .73$ for the classification of the first solution type, $\kappa = .82$ for the identification of correct solutions, and a weighted $\kappa_w = .84$ for the number of solution attempts. Using the interpretation by Landis and Koch (1977) these agreements were considered either substantial ($.60 < \kappa \leq .80$) or even "almost perfect" ($.80 < \kappa \leq 1.00$).

Preparing the coding procedure, any hints as to the experimental condition were eliminated: Transcripts from the E_5 and the E_7 condition were reduced to the mere extinction (1-solution) problem (so that all transcripts looked like those of the E_0 condition, consisting of one single task). If a subject referred to the extinction (1-solution) problem as 'task number one' (E_0 condition) or 'task number six' (E_5 condition), the number was replaced by the symbol 'X'.

From the equation-like codings the following variables were derived: (a) *number of solution attempts* within the extinction (1-solution) problem, (b) *type of first solution attempt* (correct direct solution, Einstellung solution, or other solution), (c) Einstellung repetition as indicated by the *percentage of solutions based on the original Einstellung solution*. An Einstellung solution attempt was counted whenever the beginning of a text segment contained the phrase 'B' or 'B minus'. The think-aloud protocols together with summarizing data tables of coded variables are available via the Open Science Framework (Blech et al., 2018).

Results and discussion

Behavioral data: solution frequencies and solution times

The participants quickly learned the common solution method and applied it to the first five introductory problems (see Appendix for descriptive statistics on accuracy and time needed to complete the introductory problems). We have excluded those who solved fewer than three introductory problems from the analysis on the final 1-solution problem. We considered that they did not develop a mental set to a sufficient level relative to other participants.

A number of participants also found the direct solution in the critical (2-solution) problems in the E_7 groups. This resulted in a broken mental set, which in turn led to different strategies on the final 1-solution problem compared to the participants who solved the 2-solution problems using the old established method (see Appendix, Table A2 and A3). We subsequently excluded the participants who solved the 2-solution problems using the shorter method from further analysis on the 1-solution problem.

Figure 2 demonstrates the influence of the mental set on the final extinction (1-solution) problem. The participants who experienced the Einstellung (mental set) by solving introductory E-problems were worse at finding the solution than the control participants who did not experience the Einstellung effect. However, the think-aloud protocols had no effect on the pattern of results.[2]

We ran a logistic regression where the final accuracy of the extinction (1-solution) problem was the dependent variable and thinking mode (think-aloud and silent) and experimental condition (E_0, E_5, and E_7) the two predictors. The results confirmed that there was no significant difference on the solution rate irrespective of whether the participants were thinking aloud or whether they solved the problem in silence ($b = -.44$, SE = 0.95, $z = .46$, $p = .64$, odds-ratio = .63; Nagelkerke R^2 for the whole model, .12). The analysis also confirmed that the participants were less successful in the Einstellung conditions than in the control E_0 condition ($b = -2.18$, SE = 0.82, $z = 2.66$, $p = .008$, odds-ratio = .11 for E_7 vs E_0 and $b = -1.67$, SE = 0.84, $z = 1.99$, $p = .047$, odds-ratio = .19 for E_5 vs E_0). The E-effect (i.e., the control E_0 condition superiority) was also present irrespective of the thinking mode (interactions thinking mode × condition, all $p > .42$).

Further analyses showed that with a Bayes Factor of $BF_{01} = 11.01$ the data was approximately 11 times more likely to occur under the null hypothesis of no think-aloud effect than under the alternative hypothesis of thinking aloud being reactive as to the solution accuracy of the extinction problem. Considering the factor Einstellung condition, we found $BF_{01} = 0.014$ or the inverse $BF_{10} = 70.63$. This indicates that an alternative hypothesis modeling the differences between E_7 and E_0 as well as between E_5 and E_0 was about 70 times more likely than a null hypothesis model without the effects of the Einstellung condition.

[2]Similar to previous research (Ball et al., 2015) our participants were somewhat worse in finding the solution at the beginning of the problem solving process. The difference in the first 30 seconds was, however, not significant. Cross table analysis and chi-squared tests with the two dichotomous dimensions thinking mode (aloud vs. silent) and solution (solution found within 30 seconds vs. no solution found within the first 30 seconds) with Yates correction showed no significant effects, E_7: $\chi_2(1) = 0.12$, $p = .726$; E_0: $\chi_2(1) = 2.40$, $p = .121$. In the E_5 condition the ratio of successful vs. unsuccessful problem solvers within the first 30 seconds was exactly identical (see Figure 2, middle panel).

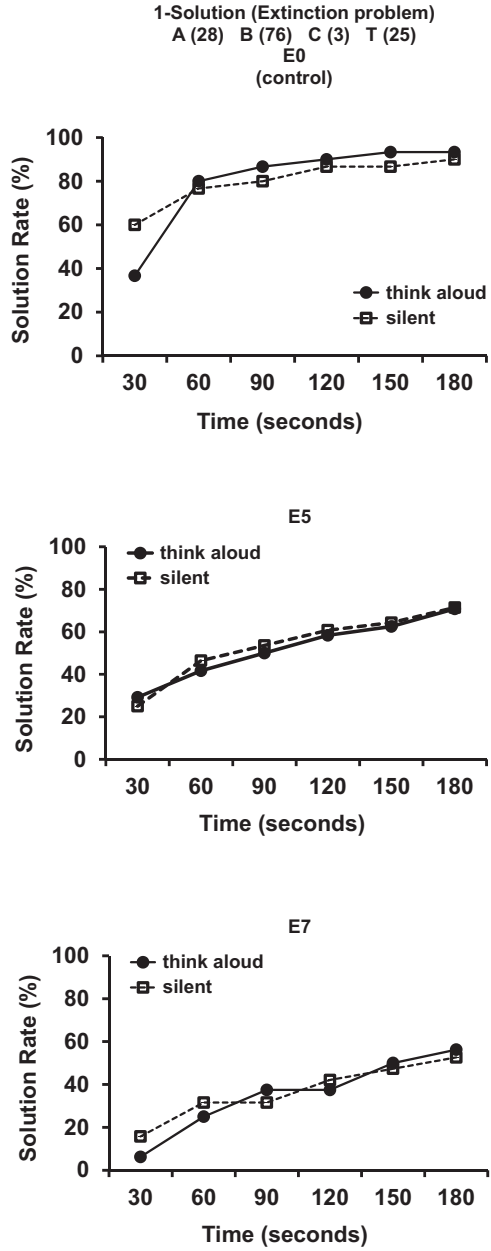

Figure 2. Solution rates in percentage on the final extinction (1-solution) problem over time (30-second groupings) in E_7 ($n_{\text{think-aloud}} = 16$; $n_{\text{silent}} = 19$), E_5 ($n_{\text{think-aloud}} = 24$; $n_{\text{silent}} = 28$), and control E_0 ($n_{\text{think-aloud}} = 30$; $n_{\text{silent}} = 30$) groups. The control group was much more successful than the two E-groups from the very beginning. Importantly, there were no significant differences between the participants who solved the problems while thinking aloud and those who did it in silence.

The same pattern of results was obtained when we considered the time needed to solve the extinction (1-solution) problem (see Appendix). Finally, the accuracy and time results indicate that more introductory problems (E_7) produced a somewhat stronger E-effect than fewer Einstellung problems (E_5; see Figures 2 and A4). However, we could not find any significant differences between E_7 and E_5 conditions.

Think-aloud data

The analysis of intermediate solution steps was based on the data of the subsample of the 80 think-aloud participants. As in the previous analysis section, within the condition E_7 we differentiated between subjects who identified at least one direct solution in the critical (2-solution) problem and those who found no direct solution at all prior to the presentation of the final extinction (1-solution) problem. Only the latter subsample was included in the following results. As an illustration of successful problem solvers and non-solvers as well as of repeated, unsuccessful solution attempts, we provide sample protocols in the appendix (Table A4), including their segmentation and coding.

Figure 3 demonstrates that the participants in the control group E_0 made fewer attempts than the participants in the other two groups which experienced the mental set problems. A one-way ANOVA confirmed the significant difference between the three groups, $F(2, 67) = 7.86$, MSE = 8.88, $p < .001$, $\eta_p^2 = .19$. Bonferroni post hoc *t*-tests confirmed that the mean number of solution attempts in the control condition ($M = 1.33$, $SD = 0.61$) was significantly lower than in the E_7 condition ($M = 2.34$, $SD = 1.01$; $t(44) = 2.98$, $p_{bonf} = .012$, $d_{Cohen} = 1.23$) and significantly lower than in the E_5 condition ($M = 2.31$, $SD = 1.44$; $t(52) = 3.58$, $p_{bonf} = .002$, $d_{Cohen} = 0.98$). The

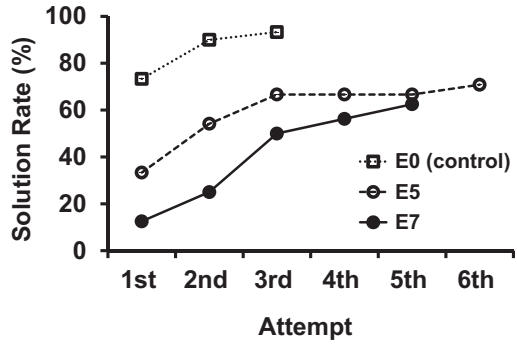

Figure 3. Solution rates in percentage on the final extinction (1-solution) problem in E_7 ($n = 16$), E_5 ($n = 24$), and control E_0 ($n = 30$) conditions over attempt order. All control group participants who solved the problem solved it by the third attempt, E_5 by the sixth attempt, and E_7 by the fifth attempt.

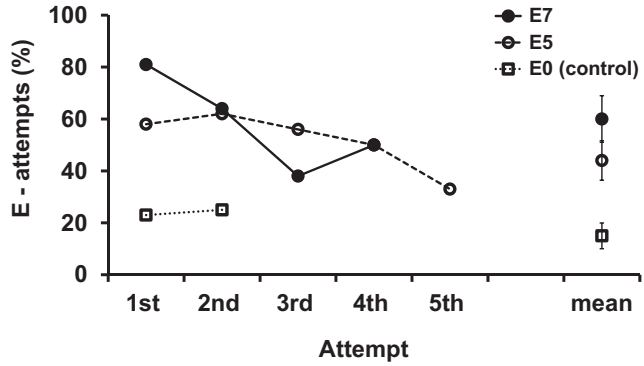

Figure 4. Percentage of the solution based on the original Einstellung solution over attempt order in E_7 ($n = 16$), E_5 ($n = 24$), and control E_0 ($n = 30$) groups. All control group participants who solved the problem solved it by the third attempt, E_5 by the sixth attempt, and E_7 by the fifth attempt. The overall average percentage of Einstellung problems, without regard to the attempt order, in the three groups is depicted at the end under "mean". Error bars represent standard error of the mean.

two Einstellung conditions did not differ significantly, $t(38) = 0.18$, $p_{bonf} = 1.000$, $d_{Cohen} = 0.05$. A Bayesian ANOVA found that the alternative hypothesis claiming an effect of the Einstellung manipulation was 43.57 times more probable than the null hypothesis, the BF_{01} being 0.023 (% error: 0.01; Bayes factors for the post hoc comparisons: E_0 vs E_5: $BF_{01} = 0.025$, % error $< .01$, E_0 vs E_7: $BF_{01} = .01$, % error $< .01$, E_5 vs E_7: $BF_{01} = 3.16$, % error $< .01$).

Figure 4 demonstrates that over half of the first attempts to solve the final extinction (1-solution) problems in E_5 and E_7 groups involved a variation on the previously acquired Einstellung solution. In contrast, only one in four of the first attempts in the control group, which did not experience the Einstellung, were attempts involving Jug B. A logistic regression using the E-solution on the first attempts confirmed that the control group tried the old E-solution less frequently compared to E_7 ($b = 2.66$, SE $= 0.77$, $z = 3.43$, $p = .001$, odds-ratio $= 14.24$; Nagelkerke $R^2 = .28$) and E_5 group ($b = 1.53$, SE $= 0.60$, $z = 2.55$, $p = .01$, odds-ratio $= 4.60$). Consistent with these findings, a Bayesian logistic regression also revealed that the data was highly (50.55 times) more likely to be observed under the alternative model including the Einstellung predictor than under the null model ($BF_{01} = 0.02$).

We also calculated the average percentage of E-solutions among the solutions for each participant. Figure 4 demonstrates that a large percentage of attempts of both Einstellung groups constituted the Einstellung solution. In contrast, only a few attempts in the control group involved using the Einstellung method of solving. A one-way ANOVA confirmed the difference

between the groups, $F(2, 67) = 11.26$, MSE = 0.11, $p < .001$, $\eta_p^2 = .25$, and Bonferroni post hoc *t*-tests indicated that the difference was between the control group on the one side, and E_5 ($t(35) = 3.3$, $p_{bonf} = .005$, $d_{Cohen} = 0.39$) and E_7 ($t(35) = 4.4$, $p_{bonf} < .001$, $d_{Cohen} = 0.53$) on the other. There were no significant differences between the two Einstellung groups, $t(23) = 1.5$, $p_{bonf} = .41$, $d_{Cohen} = 0.18$. A corresponding Bayes ANOVA indicated that an effect of the experimental condition on the percentage of attempted Einstellung solutions was 500 times more likely than a null model, $BF_{01} < 0.01$ (% error: .01; post hoc contrasts: E_0 vs E_5: $BF_{01} = 0.05$, % error < .01; E_0 vs E_7: $BF_{01} < .01$, % error < .01; E_5 vs E_7: $BF_{01} = 1.56$, % error < .01).

Figure 4 also demonstrates the perseverance of both E-groups in trying the Einstellung solution approach. Even by the third attempt, half of the attempts were related to the previously learned method that no longer worked. For participants of the Einstellung groups who did *not* find the solution of the final extinction problem at all (not included in Figure 4) the effect was still more pronounced. These subjects started with the old Einstellung solution (85% of the Einstellung variations on the first attempt) and persisted throughout to variations of the same solution, hardly trying different paths (73%, 75%, 66% of the Einstellung variation for the second, third, and fourth attempt).

General discussion

Given the amount of research on the Einstellung effect, it is not surprising that we have replicated the original water jug results. When people learn a way of doing things, it becomes difficult to find another way of dealing with a situation where the old method is not applicable. The E-effect in Luchins' water jug paradigm remains one of the most replicable phenomena (Luchins & Luchins, 1994).

Non-reactivity and potentials of Think-Aloud protocols in Einstellung research

Our study adds to the current knowledge in two ways. First, this is, to our knowledge, the only study that directly tackled the question of the think-aloud technique's suitability in the water jug paradigm. Our results demonstrate that the think-aloud technique is a valid tool for tracking the activation of representations in the water jug paradigm. In all three groups, there was no indication in the behavioral measures that thinking aloud influences the very processes it is meant to measure. Neither the quality of the final solutions nor the solution times were affected by the thinking

mode, even when we used Bayesian analysis instead of the classical inferential analyses. The results are consistent with the findings of the meta-analysis (Fox et al., 2011) where concurrent thinking aloud was not found to interfere with the ongoing cognitive processes – unlike other seemingly similar reporting techniques, which involve a high degree of self-reflection and self-monitoring, e.g., self-explanations (Neuman & Schwarz, 1998). The results are also in alignment with the study by Ball and Stevens (2009) in which thinking aloud did not impair the performance on verbal insight problems (compound remote associates). In a similar vein, Fleck and Weisberg (2004) could find no difference between a think-aloud and a silent group on the famous candlestick problem (Duncker, 1945). Given that task complexity played a role with thinking aloud being beneficial in solving the more complex problems, one might assume that the moderately complex water jug tasks fell in the intermediate range of neither positive nor negative effects of overt verbalization.

Our present finding was, however, not in line with the literature on verbal overshadowing (Schooler et al., 1993), the special-process theory (Seifert et al., 1995) and the results by Ball et al. (2015) on thinking-aloud effects in visual insight problems. One possible reason for the different findings is the nature of the problems used. Like a good number of other insight problems, the problems in Ball et al. were visually based. In contrast, the water jug task is a verbal problem, as in Ball and Stevens (2009). It can be dismantled into stepwise components with gradual arithmetic approaches towards the final solution as formalized in the process model by Atwood and colleagues (Atwood, Masson, & Polson, 1980; Atwood & Polson, 1976), who emphasize the impact of planning activities in the water-jug problem (see also Delaney et al., 2004). Hence it is possible that such step-by-step solutions are not impaired or may even benefit from thinking aloud. The verbalization can guide cognitive processes, especially when a problem solver – prompted or unasked – includes elaborate self-explanations (Chi et al., 1989; Neuman & Schwarz, 1998; Redifer et al., 2016). The assumed negative and positive effects of problem-related speech could intermix, leveling each other out in the water jug task.

The think-aloud analysis also limits alternative explanations. For example, given that the participants were always using all three jugs in the introductory problems, they may have explicitly assumed that they have to use all three jugs when solving the problems. The fixation on using all three jugs is possibly a contributing factor to the overall E-effect that future research needs to account for. We do, however, believe that it cannot possibly be the main one. Our think-aloud protocols show that 80% of the first attempts start from Jug B. Once participants start solving the extinction

problem (1-solution) with Jug B, it is impossible to solve the problem no matter whether one uses two or three jugs.

Evidence for mutual strengthening of activated memory and allocated attention

Secondly, and most importantly, think-aloud protocols provide evidence about the workings of the E-effect. We demonstrated that the E-effect is rooted in sustained activation of inappropriate elements. It is not surprising that the majority of the first attempts are related to the previous E-solution (Figure 3). After all, that old E-solution had been working well in the previous E-problems. What is surprising is that a good number of participants kept trying to make the old method work despite its repeated failure. Not only were most of the second attempts variations on the old solution, but even the third and the fourth attempts were dominated by versions of the old solving method (Figure 4). Similar to the mechanism found in the eye tracking studies on the E-effect in experts (Bilalić et al., 2008b), sustained activation of the ineffective solution scheme in working memory was the main culprit behind participants' remarkable blindness. Abandoning one solution attempt in order to try another costs time and effort in case of procedural (e.g., Woltz et al., 2000) or declarative representations. Analogous costs and mechanisms have been proposed for switching the focus of attention between objects held in working memory (see Gade, Souza, Druey, & Oberauer, 2017; Oberauer, Souza, Druey, & Gade, 2013).

It is important to stress that the uncovered E-mechanism is not only limited to the water-jug paradigm. It has already been shown that the same mechanism, where activated memory biases intake of perceptual clues by driving the attentional resources towards elements related to the activated elements in memory, explains the rare mistakes of experts (Bilalić et al., 2008a, 2008b; Sheridan & Reingold, 2013). The E-mechanism also bears resemblance to a number of seemingly unrelated phenomena (Bilalić et al., 2010; Bilalić & McLeod, 2014). One of the explanations for impasse in insight problem solving, for example, assumes that fixed mindset effects arise from the unwanted activation of a wrong solution (or a misleading cue), which blocks working memory and diverts attention from a more successful approach (e.g., the activation hypothesis by Smith, 1995b). Despite constant failures to find the solution to insight problems, it is almost impossible for people not to think of the same method when they face the same problem again – a phenomenon akin to the one the participants in this study experienced.

Another similar phenomenon is design fixation, which refers to adherence to a particular set of idea or concepts during the design process (Jansson &

Smith, 1991). In a typical paradigm, the group of designers who are presented with a faulty example solution to a (design) problem generate solutions that regularly involve the suboptimal features from the presented example solution. In contrast, the group of designers who were only given the design problem without the faulty solution creates solutions mostly without the inadequate features. The inability of the primed group to set aside the activated concepts thought the solution, even if it is inadequate, is not unlike the E-phenomena and its associated E-mechanism (Crilly & Cardoso, 2017). The similarity has been acknowledged by the design fixation community, which has recently proposed an alternative way of investigating design fixation employing a mental set like paradigm (Neroni, Vasconcelos, & Crilly, 2017).

People often do not change a point of view, even when they are motivated for the change and examine new evidence with a seemingly open mind (Lord, Ross, & Lepper, 1979). One of the reasons may lie in an Einstellung-like inability to consider situational aspects that are unrelated to their initial opinion. The activated idea biases their attention towards evidence that is related to the idea, which in turn reinforces the initial opinion even further. More importantly, given that the person has exerted effort to understand the evidence, the whole process leaves the person with a false impression of being open-minded.

The prevalence of an Einstellung-like mechanism in cognition makes it important to find methods for reducing or eliminating its adverse effects. One way to devise beneficial strategies for combating any phenomenon is to understand the working behind it. For example, we know that drawing attention to the inappropriate elements, even if it is to say that they should think of something else, is not helpful. Arguably the most prominent example of such a strategy is the famous study where the participants were instructed not to think about a white bear (Wegner, Schneider, Carter, & White, 1987; Wegner & Schneider, 2003). The study illustrates that actively trying to distract one's attention from a particular object can result in the opposite effect – the activation will be strengthened rather than weakened (Giuliano & Wicha, 2010). Instead, future studies on the E-effect may directly draw attention to the crucial elements and away from the initially activated elements (e.g., Grant & Spivey, 2003; Thomas & Lleras, 2008). Another promising avenue may involve weakening the initial memory activation through unrelated activity or even a simple break (Sio & Ormerod, 2009).

Conclusion

Our study underlines the usefulness of think-aloud technique in tracking cognitive processes. Not only did think-aloud not influence the problem solving process itself, but it also enabled us to rule out important

alternative explanations (e.g., the three-jug hypothesis) that would be difficult to do with other techniques. Most importantly, it allowed us to elicit the pernicious nature of the E-effect. Despite repeatedly failing to solve the problem using the previously successful method, the participants kept applying variations of the same method. The uncovered synergy between the initially activated memory and subsequent allocated attention and perception dependent on it should be used in future research looking into finding ways to counter the Einstellung mechanism.

Acknowledgments

We would like to thank the students taking the project course 'Empirisch-experimentelles Praktikum' in the summer term of 2017 at the FernUniversität in Hagen. We also thank Annika Fischer and Christina Luchsinger for coding the transcripts and Christina Luchsinger for translating test material provided on the Open Science Framework as well as Peter McLeod and Linden Ball for their comments on the manuscript. Matthew Bladen's contribution to preparing the article for publication is also greatly appreciated.

Disclosure statement

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

References

Anderson, J. R. (2013). *The architecture of cognition*. New York: Psychology Press.

Atwood, M. E., Masson, M. E., & Polson, P. G. (1980). Further explorations with a process model for water jug problems. *Memory & Cognition*, *8*(2), 182–192. doi:10.3758/BF03213422

Atwood, M. E., & Polson, P. G. (1976). A process model for water jug problems. *Cognitive Psychology*, *8*(2), 191–216.

Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford: Oxford University Press.

Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173. doi:10.1037/0033-295X.105.1.158

Ball, L. J., Marsh, J. E., Litchfield, D., Cook, R. L., & Booth, N. (2015). When distraction helps: Evidence that concurrent articulation and irrelevant speech can facilitate insight problem solving. *Thinking & Reasoning*, *21*, 76–96. doi:10.1080/13546783.2014.934399

Ball, L. J., & Stevens, A. (2009). Evidence for a verbally-based analytic component to insight problem solving. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 1060–1065). Austin, TX: Cognitive Science Society.

Bilalić, M., & McLeod, P. (2014). Why good thoughts block better ones. *Scientific American*, *310*(3), 74–79. doi:10.1038/scientificamerican0314-74

Bilalić, M., McLeod, P., & Gobet, F. (2008a). Expert and 'novice' problem solving strategies in chess: Sixty years of citing de Groot (1946). *Thinking & Reasoning*, *14*, 395–408. doi:10.1080/13546780802265547

Bilalić, M., McLeod, P., & Gobet, F. (2008b). Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition*, *108*(3), 652–661. doi:10.1016/j.cognition.2008.05.005

Bilalić, M., McLeod, P., & Gobet, F. (2010). The mechanism of the Einstellung (set) effect: A pervasive source of cognitive bias. *Current Directions in Psychological Science*, *19*(2), 111–115.

Blech, C., Gaschler, R., & Bilalić, M. (2018). Why do people fail to see simple solutions? Using think-aloud protocols to uncover the mechanism behind the Einstellung (mental set) effect. *Open Science Framework*. Retrieved from https://osf.io/78vua/.

Ceruzzi, P. E. (2003). *A history of modern computing* (2nd ed.). Cambridge, MA: MIT Press.

Chen, Z., & Mo, L. (2004). Schema induction in problem solving: A multidimensional analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 583–600.

Chi, M., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. doi:10.1207/s15516709cog1302_1

Crilly, N., & Cardoso, C. (2017). Where next for research on fixation, inspiration and creativity in design? *Design Studies*, *50*, 1–38. doi:10.1016/j.destud.2017.02.001

Crosskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, *78*, 775–780.

Delaney, P. F., Ericsson, K. A., & Knowles, M. E. (2004). Immediate and sustained effects of planning in a problem-solving task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1219–1234.

Dresing, T., Pehl, T., & Schmieder, C. (2015). *Manual (on) transcription. Transcription conventions, software guides and practical hints for qualitative researchers*. 3rd English edition. Retrieved from http://www.audiotranskription.de/english/

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, *58*(5), i–113. doi:10.1037/h0093599

Ellis, J. J., & Reingold, E. M. (2014). The Einstellung effect in anagram problem solving: Evidence from eye movements. *Frontiers in Psychology*, *5*, 679.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J., Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 223–241). New York, NY: Cambridge University Press

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: The MIT Press.

Fleck, J. I., & Weisberg, R. W. (2004). The use of verbal protocols as data: An analysis of insight in the candle problem. *Memory & Cognition*, *32*(6), 990–1006. doi:10.3758/bf03196876

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. doi:10.1037/a0021663

Gade, M., Souza, A. S., Druey, M. D., & Oberauer, K. (2017). Analogous selection processes in declarative and procedural working memory: N-2 list-repetition and task-repetition costs. *Memory & Cognition*, *45*, 26–39. doi:10.3758/s13421-016-0645-4

Giuliano, R. J., & Wicha, N. Y. Y. (2010). Why the white bear is still there: Electrophysiological evidence for ironic semantic activation during thought suppression. *Brain Research*, *1316*, 62–74. doi:10.1016/j.brainres.2009.12.041

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, *14*(5), 462–466. doi:10.1111/1467-9280.02454

Greeno, J. G., Magone, M. E., & Chaiklin, S. (1979). Theory of constructions and set in problem solving. *Memory & Cognition*, *7*(6), 445–461. doi:10.3758/BF03198261

Jansson, D. G., & Smith, S. M. (1991). Design fixation. *Design Studies*, *12*(1), 3–11. doi:10.1016/0142-694X(91)90003-F

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. doi:10.1037/0022-3514.37.11.2098

Lovett, M. C., & Anderson, J. R. (1996). History of success and current context in problem solving: Combined influences on operator selection. *Cognitive Psychology*, *31*(2), 168–217. doi:10.1006/cogp.1996.0016

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs*, *54*(6), i–95. doi:10.1037/h0093502

Luchins, A. S., & Luchins, E. H. (1994). The water jar experiments and Einstellung effects: II. Gestalt psychology and past experience. *Gestalt Theory*, *16*(4), 205–259.

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

Neroni, M., Vasconcelos, L., & Crilly, N. (2017). Computer-based "mental set" tasks: An alternative approach to studying design fixation. *Journal of Mechanical Design*, *139*(7), 071102. doi:10.1115/1.4036562

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Neuman, Y., & Schwarz, B. (1998). Is self-explanation while solving problems helpful? The case of analogical problem-solving. *British Journal of Educational Psychology*, *68*(1), 15–24. doi:10.1111/j.2044-8279.1998.tb01271.x

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. doi:10.1037/1089-2680.2.2.175

Nijstad, B. A., & Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review*, *10*(3), 186–213. doi:10.1207/s15327957pspr1003_1

Oberauer, K., Souza, A. S., Druey, M. D., & Gade, M. (2013). Analogous mechanisms of selection and updating in declarative and procedural working memory: Experiments and a computational model. *Cognitive Psychology*, *66*(2), 157–211. doi:10.1016/j.cogpsych.2012.11.001

Öllinger, M., Jones, G., & Knoblich, G. (2008). Investigating the effect of mental set on insight problem solving. *Experimental Psychology*, *55*(4), 269–282. doi:10.1027/1618-3169.55.4.269

Penney, C. G., Godsell, A., Scott, A., & Balsom, R. (2004). Problem variables that promote incubation effects. *The Journal of Creative Behavior*, *38*(1), 35–55. doi:10.1002/j.2162-6057.2004.tb01230.x

Redifer, J. L., Therriault, D. J., Lee, C. S., & Schroeder, A. N. (2016). Working memory capacity and self-explanation strategy use provide additive problem-solving benefits. *Applied Cognitive Psychology*, *30*(3), 420–429. doi:10.1002/acp.3219

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*(2), 166–183. doi:10.1037/0096-3445.122.2.166

Schultz, P. W., & Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genetic, Social, and General Psychology Monographs*, *128*(2), 165–207.

Seifert, C. M., Meyer, D. E., Davidson, N., Patalano, A. L., & Yaniv, I. (1995). Demystification of cognitive insight: Opportunistic assimilation and the prepared-mind perspective. In R. J. Sternberg & J. E. Davidson (Eds.) *The nature of insight* (pp. 65–124). Cambridge, MA: MIT Press.

Sheridan, H., & Reingold, E. M. (2013). The mechanisms and boundary conditions of the Einstellung effect in chess: Evidence from eye movements. *PloS One*, *8*(10), e75796. doi:10.1371/journal.pone.0075796

Sio, U. N., & Ormerod, T. C. (2009). Does incubation enhance problem solving? A meta-analytic review. *Psychological Bulletin*, *135*(1), 94–120. doi:10.1037/a0014212

Smith, S. M. (1995a). Fixation, incubation, and insight in memory and creative thinking. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 135–156). Cambridge, MA, US: The MIT Press.

Smith, S. M. (1995b). Getting into and out of mental ruts: A theory of fixation, incubation, and insight. In R. J. Sternberg & J. E. Davidson (Eds.) *The nature of insight* (pp. 229–251). Cambridge, MA: MIT Press.

Thomas, L. E., & Lleras, A. (2008). Moving thought: Directed movement guides insight in problem solving. *Journal of Vision*, *8*(6), 1053–1053. doi:10.1167/8.6.1053

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, *53*(1), 5–13. doi:10.1037//0022-3514.53.1.5

Wegner, D. M., & Schneider, D. J. (2003). The white bear story. *Psychological Inquiry*, *14*(3), 326–329. doi:10.1037/0022-3514.53.1.5

Woltz, D. J., Gardner, M. K., & Bell, B. G. (2000). Negative transfer errors in sequential cognitive skills: Strong-but-wrong sequence application. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 601–625. doi:10.1037//0278-7393.26.3.601

Appendix

Additional analyses and materials

Introductory (Einstellung) problems

The eight water jug problems we used in the experiment (variations from Luchins, 1942) are presented in Table A. The first five problems are called Einstellung or introductory problems because they are inducing mental set in the participants – all can be solved using the same B – A – 2C method. Problems 6 and 7 are called critical or 2-solution problems because they can be solved by an old, long solution (B – A – 2C) or a new, short solution (A – C and A + C, respectively). Problem 8 is

Table A1. Einstellung, critical and extinction problems in the water-jug paradigm.

		Capacity of the jugs			
Problem		Jug A	Jug B	Jug C	Target
0-a	*warm-up 1*	*29*	*3*	–	*20*
0-b	*warm-up 2*	*15*	*2*	–	*13*
1	*Einstellung* 1	21	127	3	100
2	*Einstellung* 2	20	59	4	31
3	*Einstellung* 3	24	47	7	9
4	*Einstellung* 4	33	76	15	13
5	*Einstellung* 5	18	43	10	5
6	Critical 1 (2-solution)	23	49	3	20
7	Critical 2 (2-solution)	15	39	3	18
8	**Extinction (1-solution)**	**28**	**76**	**3**	**25**

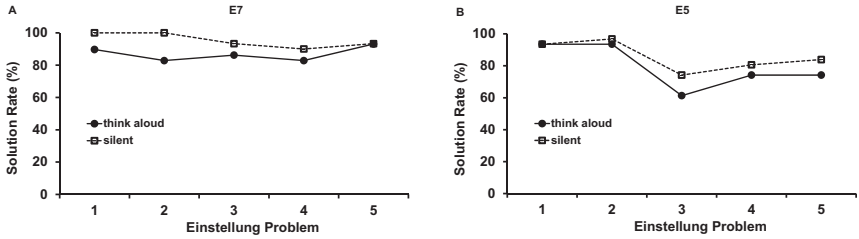

Figure A1. The solution accuracy on the Einstellung (introductory) problems.

called the extinction or 1-solution problem because it can be solved only by the new shorter solution (A – C).

Group E_7 is so called because it was shown all seven problems before solving the final extinction (1-solution) problem. Group E_5 solved only the first five introductory problems without the critical problems, while the control E_0 solved only the extinction (1-solution) problem. The solution accuracy of E_7 and E_5 groups in the introductory problems, depending on the thinking mode, can be seen in Figure A1. Think-aloud participants and silent thinkers did not differ significantly in their solution rates as averaged over the five Einstellung problems, neither in the E_7 group, $t(54) = -0.75$, $p = .46$, $d_{Cohen} = 0.20$, nor in the E_5 group, $t(50) = 0.03$, $p = .97$, $d_{Cohen} = 0.01$. In line with this conclusion, Bayesian independent samples *t*-Tests found Bayes Factors of $BF_{01} = 2.93$ (E_7; error %: .01) and $BF_{01} = 3.58$ (E_5; error %: .02), being little stronger in favor of the null hypothesis of no reactivity than of the alternative hypothesis.

The time needed for the two groups to find the correct solution in the introductory problems can be seen in Figure A2. Although in the E_7 condition for the problems 1 ($p = .03$) and 2 ($p = .02$) the time need was higher for subjects thinking aloud, averaged over the five Einstellung problems, think-aloud participants did not differ significantly in their solution time (E_7: $t(54) = 1.40$, $p = .17$, $d_{Cohen} = 0.38$; E_5: $t(50) = 0.62$, $p = .54$, $d_{Cohen} = 0.19$) as could also be seen from Bayesian independent *t*-tests. With Bayes factors of $BF_{01} = 1.64$ (E_7; % error: .01) and $BF_{01} = 3.06$ (E_5; % error: .02) there was neither strong evidence of the null hypothesis nor of the alternative hypothesis claiming a reactive effect of thinking aloud on solution time.

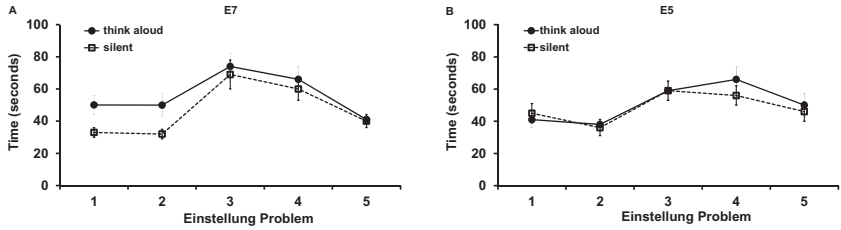

Figure A2. Time need to correctly solve the Einstellung (introductory) problems.

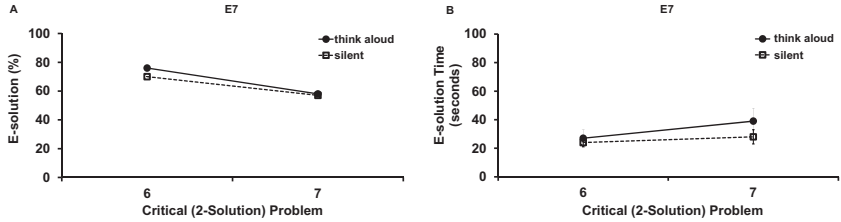

Figure A3. (A) Percentage of participants finding the old Einstellung solution in the critical (2-solution) problems. (B) Time needed to find the Einstellung solution in the critical (2-solution) problems (error bars represent 1 SEM).

2-solution (critical) problems

Figure A3 shows that the majority of the participants opted for the old Einstellung solution in the critical problems. Logistic regression with thinking mode as a predictor showed that there were no significant differences between thinking aloud and thinking in silence, neither for problem 6 ($b = .36$, SE $= .61$, $z = -0.15$, $p = .56$, odds-ratio $= 1.43$; Nagelkerke $R^2 = .01$) nor for problem 7 ($b = -.20$., SE $= .55$, $z = -0.55$, $p = .71$, odds-ratio $= 0.82$; Nagelkerke $R^2 < .01$). Similarly, the Bayes factors ($BF_{01} = 6.31$ for problem 6; $BF_{01} = 6.99$ for problem 7) pointed out that the occurrence of the data was at least six times more likely under the null hypothesis of no think-aloud effect than under the alternative hypothesis. The time needed to find the familiar solution was also rather quick and, averaged over problem 6 and 7, not significantly affected by the thinking mode, $t(54) = 1.37$, $p = .18$, $d_{Cohen} = 0.37$. A Bayes factor of $BF_{01} = 1.71$ (error %: .01) yielded anecdotal support for the null hypothesis of no thinking aloud effect.

2-solution (critical) problem solvers on the 1-solution (extinction) problem

The implementation of the 2-solution problems replicates the original Luchins paradigm and is well suited to demonstrating the preference for routine solutions, but is suboptimal when it comes to investigating the think-aloud correlates of a mental set. Comparisons between subjects of the E_7 condition who used the Einstellung solution in both critical trials and participants who found the simple solution in at least one of the critical trials (problem 6 and problem 7), suggested that for the latter group the mental set was no longer intact. Contrasted with the

Table A2. Comparison of critical problem solvers (solvers) and non-solvers regarding their solution time for the extinction problem, their number of solution attempts, and the relative frequency of Einstellung solution attempts as identified in the think-aloud protocols.

	Solvers		Non-solvers			
Dependent variable	*M*	*SD*	*M*	*SD*	*t*-Test	BF_{01}
Solution time (seconds) for extinction problem[a]	12.05	6.13	99.11	10.39	$t(35.12) = 8.31$, $p < .001$, $d_{Cohen} = 1.99$	<0.01
Number of solution attempts[b]	1.10	0.32	2.31	1.08	$t(18.82) = 4.22$, $p < .001$, $d_{Cohen} = 1.53$	0.06
Percentage of Einstellung solution attempts[b]	0.05	0.16	0.60	0.36	$t(22.18) = 5.39$, $p < .001$, $d_{Cohen} = 2.00$	<0.01

Note. [a]Behavioral data: 21 critical problem solvers, 35 non-solvers.
[b]Think-aloud data: 10 critical problem solvers, 16 non-solvers.

Table A3. Logistic regressions comparing critical problem solvers and non-solvers of the Einstellung E_7 condition with respect to their problem-solving outcome and approaches in the extinction trial. Solution accuracy was coded with a positive outcome score (1) if a problem solver named the correct Einstellung solution within 180 seconds, otherwise it was coded as zero (0). Einstellung as a first solution attempt was coded from the verbal protocols.

	Coefficients for dummy *critical problem solver*					Model estimates	
Dependent variable	*b*	SE	z	*p*	odds-ratio	Nagelkerke R^2	BF_{01}
Solution accuracy[a]	19.39	0.40	48.07	<.001	264,700,000	.41	<.01
Einstellung as first solution attempt[b]	−3.66	1.23	−2.97	.003	0.03	.56	<.01

Note. [a]Behavioral data: 21 critical problem solvers, 35 non-solvers.
[b]Think-aloud data: 10 critical problem solvers, 16 non-solvers.

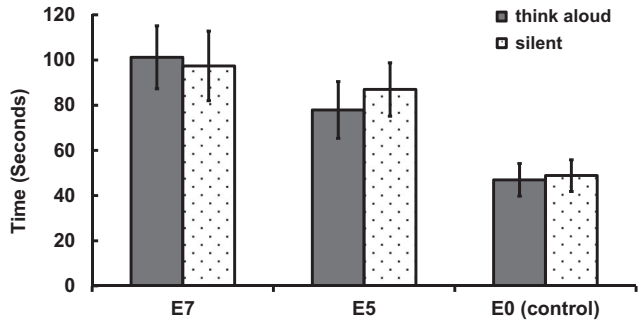

Figure A4. Time need to solve the final extinction (1-solution) problem depending on the group and thinking mode (error bars represent 1 SEM).

non-solvers, the so-called critical problem solvers were significantly faster and more accurate in solving the subsequent extinction problem; they took fewer solution attempts, and their attempts – both the first attempt as well as overall – were less often classified as Einstellung solutions. In short, their solution patterns compiled in Tables A2 and A3 resembled the E_0 subjects much more than either the E_5

Table A4. Sample protocols and their segmentation into solution attempts for participants who needed several attempts before either finding the direct solution (#19, #21) or not finding the solution within the time frame of 180 seconds (#104, #122). The passages marked with suspension points ("…") in brackets mark short breaks in verbalization of one up to a few seconds. The analysis refers to the final extinction (1-solution) problem in which the parameters were as follows: jug A = 28, jug B = 76, jug C = 3, target jug = 25.

German original	English translation	Attempt segmentation
Subject: #104, condition E_5, non-solver		
25 Zielmenge. Krug B hat 76. (…) 76, 76, 56, 56. Jetzt hänge ich gerade. 56, 50, 48. Hah. 48, 45. Mhhh. Mhhh. (…) Habe ich mich zu sehr darauf verlassen, dass das ein und das gleiche ist.	25 is the target amount. Jug B has 76 (…) 76, 76, 56, 56. I'm stuck now. 56, 50, 48. Hah. 48, 45. Mhhh. Mhhh. (…) I've put too much faith in the same method.	Attempt 1: B – A – C
76 (…) Einmal 28 sind 56. Das wären dann 76 minus 56 wären 20. Und minus 3. Nein. 56 Nein. Das geht ja auch nicht. (..) Schade.	76 (…) once 28 makes 56. Then 76 minus 56 would be 20. And minus 3. No. 56. No, that doesn't work either (…) pity.	Attempt 2: B – 56 – C
76 minus zweimal Krug A bleiben mir nur 20. Wenn ich dann Krug C dazu tue, dann hab ich 23. Da komme ich nicht hin. Das ist ja ein Schmarrn.	76 minus twice jug A, makes 20 left. When I add jug C, I get 23. I don't get it. That's nonsense.	Attempt 3: B – 2A + C
Also 76 minus 28, 56, 48. Irgendwo habe ich einen Fehler gemacht. (…) 48 minus 3 sind 45. Das haut nicht hin. (…) Mhhh. Wo habe ich da jetzt den Wurm drin? (…) Wenn ich es so herum mache, dann fehlen mir 5. Die habe ich aber nirgendswo. 56. Ich lach mich tot. (…)	So 76 minus 28, 56, 48. Somewhere I made a mistake. (…) 48 minus 3 yields 45. That does not make sense. (…) Mhhh. Now, where's the mistake? (…) If I try it that way, I miss out 5. And I don't have them. 56. I laugh my head off.	Attempt 4: B – A - C
Ah. Ich brauch ja nicht nur abziehen. Ich kann ja auch dazurechnen. 48 (…) 48 (…) 56, 48. Ich komme nicht drauf. Das kann nicht sein. Aber …	Ah. I don't have to subtract. I can add, too. 48 (…) 48 (…) 56, 48. I don't get it. That's not possible. But …	Attempt 5: 56
Subject: #104, condition E_5, solver		
76 … 76, also 28 plus 28 sind 76, nein, sind 56 (…) das ist zu wenig, also, 76 minus 28 sind 48 (…) minus 28 sind 20 (…), geht auch nicht.	76 … 76, so 28 plus 28, that is 76, no, 56 (…) that is too low, so 76 minus 28 brings 48 (…) minus 28 makes 20 (…), won't work, either.	Attempt 1: A + A … B – A – A
76 minus 3 sind 73, minus 28 sind (..) 48. 48 minus 28 wären (…) ne, das geht auch nicht.	76 minus 3 is 73, minus 28 is (..) 48. 48 minus 28 would be (…) no, does not work, either.	Attempt 2: B – C- A - A
Ähm, ja, also nochmal. 76 minus 28, dann bin ich bei (…) 48 (…) minus 28, das ist zu wenig. Also minus 3, dann bin ich bei 45 (…) wie komm ich denn da auf die 25? Hmm, (…)	Well, yes, once again. 76 minus 28, then I get (…) 48 (…) minus 28, that is too low. So minus 3, gets me to 45 (…) how can I get to 25? Hmm, (…)	Attempt 3: B – A – A – C
28 plus 48, 28, sind 56, das kommt aber auch nicht auf (…)	28 plus 48, 28, that's 56, but that doesn't hold, either (…)	Attempt 4: A + 48

(*continued*)

Table A4. Continued.

German original	English translation	Attempt segmentation
also muss ich (…) 10 mal 3 sind 30, ich überlege jetzt wie ich es kombinieren kann … aber das ähm, ohne aufzuschreiben ist das schwierig finde ich. Hmm. Also wenn ich jetzt 76 minus 3, minus 3 bin ich bei 70 (…) Also wenn ich den Krug C, kann ich 30 mal füllen, dann wäre ich bei 30, 20 mal füllen wäre ich bei 60. Kommt ja auch nicht.	Hence I need to (…) 10 times 3 yields 30, now I'm wondering how to combine it … but that is, er, without taking it down, it's difficult, I find. Hmm. So when I try 76 minus 3 now, I get to 70 (…) When I fill jug C 30 times, then I would have 30, 20 times filling would be 60. Does not work.	Attempt 5: (10*C) * (-1) + B – C – C
(…). 76 minus 25, dann bin ich bei 51 (…) geht aber auch nicht.	(…). 76 minus 25, then I have 51 (…) won't work, either.	Attempt 6: B – T
28. Ach ich bin vielleicht doof. Ja, also die 28 minus die 3 wären ja auch 25 gewesen, aber ich habe es mir unheimlich schwer gemacht. Hab es zum Schluss erst gesehen, ja.	28. What am I stupid. Yes, so the 28 minus the 3, that's it, but I did it the hard way. Didn't see it up to the end, indeed.	Attempt 7: A – C
Subject: #122, condition E_5, non-solver		
Äh. 76 minus 28 (…) Oh Gott, ich (…) 78. 76 minus 20 ist 56. (…) Sind 48. 48; minus 20 sind achtund/ äh. (.) 28. Nein. Geht nicht. (…)	Er, 76 minus 28 (…) Oh my God, I (…) 78. 76 minus 20, that is 56. (…) brings 48. 48; minus 20 results in eight … er 28. No. Won't work (…)	Attempt 1: B – A
76, minus 28. Ähm. (…) 56, minus 8 sind 48. (…) 48. (…) Äh. 48, äh, minus 3 sind 45, minus/ (..) 3 sind (..) 42. Minus 3 (..) sind (…) Ich kann mir das nicht vorstellen. (…) Äh. Ich kriege die nicht heraus. (…)	76, minus 28. Ugh. (…) 56, minus 8, that is 48. (…) 48. (…) Er. 48, er, minus 3 brings 45, minus (..) 3 makes (..) 42. Minus 3 (..) that is (…) I cannot understand this. (…) Hmm. I cannot figure it out.	Attempt 2: B – A
Ah. (…) 76, minus 28 sind 76, minus 28 (…) 48. Minus 6. Minus 6 sind … Äh. (…) Kann ich mir das aufschreiben? *Versuchsleitung: Nein, leider nicht.* 76, minus 28 (…) Minus 20 sind 56. Minus 8 sind 48. Achtund…	Ah. (…) 76, minus 28 brings 76, minus 28 (…) 48. Minus 6. Minus 6 that is … Ugh. (…) Can I take it down? *Experimenter: No, I'm afraid not.* 76, minus 28 (…) Minus 20 that is 56. Minus 8 brings 48. Eight…	Attempt 3: B – A
Subject: #21, condition E_7, solver		
76 minus 28 sind äh 56, sind 48, sind 48. Minus 48 minus 3 sind 45. (…) Fehlen noch (…) 20 (…).	76 minus 28 that is, er, 56, that is 48, is 48. Minus 48 minus 3 that is 45. (…) still needs (…) 20 (…).	Attempt 1: B – A – C
Nochmal neu! 76. 2 mal 28 sind (..) ähm. (…) Bringt mir das was? Also 76 minus 28. 76 minus 28 sind 6 und (…) viert… Hä, bin ich blöd? (…) Sind 56 sind äh 48. (…) Ähm, sind (…)	Once again, anew! 76. 2 times 28 brings (..) hmm. (…) Will it work? So 76 minus 28. 76 minus 28 brings 6 and (…) fort… Well, am I stupid? (…) That's 56, is, er 48. (…) Er, that is (…)	Attempt 2: 2A, B – A
Ähm anders. Ähm 28 (…) und (…) Ach bin ich blöd! Bin ich blöd! Krug A minus Krug C. Meine Herren. A minus C. Meine Herren.	Hmm, differently. Hmm 28 (…) and (…) Oh, I'm silly! How am I silly! Jug A minus Jug C. My goodness. A minus C. My goodness.	Attempt 3: A – C

or the remainder of the E_7 participants. A possible explanation for why the critical solvers mental set was not well established was the observation that in the first five Einstellung trials, the later critical solvers tended to find the correct Einstellung solution less often than non-solvers of the E_7 condition, $t(24.56) = 2.66$, $p = .014$, $d_{Cohen} = 0.80$, $BF_{01} = 0.07$, while the mean solution time was not significantly different, $t(54) = -1.29$, $p = 0.201$, $d_{Cohen} = -0.36$, $BF_{01} = 1.81$.

When we checked the time needed to solve the final extinction (1-solution) problem (Figure A4), we found the same pattern of results as the one concerning accuracy in the main text. A 3×2 ANOVA on the time need to solve the extinction (1-solution) problem with thinking mode (think aloud, silent) and condition (E_0, E_5, and E_7) as between factors produced no significant difference between silent and think-aloud mode across the three conditions, $F(1, 141) = 0.06$, MSE = 194.2, $p = .80$, $\eta_p^2 < .001$. The thinking mode did not influence the difference between conditions (interaction thinking mode × condition: $F(2, 141) = .1$, MSE = 447.8, $p = .87$, $\eta_p^2 = .002$). However, the Einstellung effect itself was significant (main effect of condition: $F(2, 141) = 10.69$, MSE = 33161.6, $p < .001$, $\eta_p^2 = .131$) as the control E_0 condition yielded significantly quicker solutions than E_5 ($p_{bonf} = .004$) and E_7 conditions ($p_{bonf} < .001$). There were no differences between the two E conditions (E_5 vs E_7: $p_{bonf} = .511$). An additional Bayesian ANOVA provided comparable results: for the thinking mode with $BF_{01} = 5.06$ (error %: < 0.01) the odds were about five times higher in favor of the null hypothesis; for the factor condition with $BF_{01} < 0.01$ or $BF_{10} = 574.97$ respectively, the alternative hypothesis was clearly more likely. The Bayes factors for the post hoc comparisons were: $BF_{01} = 0.03$ for E_0 vs E_5, $BF_{01} < 0.01$ for E_0 vs E_7, and $BF_{01} = 2.32$ for E_5 vs E_7.

Sample protocols, segmentation, and coding scheme

As an example of a typical problem solver from the E_0 control group consider the following translated short protocol: "*So … I see three jugs here. Oh no, big numbers, 28, 76 and 3 and the target amount is. And since I have a water fountain, I'd first fill in jug A, the jug then put it in jug C so that 3 liters is subtracted, so that 28 minus 3 is 25*." The solution was classified as consisting of one single attempt coded as A (28) – C (3), the direct solution.

Longer protocols showed sequences of several solution attempts, repetitions of the Einstellung solution (attempts beginning with "B" and especially "B minus"), sometimes an insight-like finding of the direct solution towards the end, while others were running out of time without identifying the A – C solution. The four sample protocols in Table A4 illustrate their cognitive impasses as well as the segmentation and coding procedure.