CREATIVITY RESEARCH JOURNAL
https://doi.org/10.1080/10400419.2023.2200597

The Impact of Adding a Fourth Item to the Traditional 3-Item Remote Associates Test

Jose A. Diaz[a,b], Steven M. Nelson[a,c], A. Alexander Beaujean[a], Adam E. Green[d], and Michael K. Scullin[a]

[a]Department of Psychology and Neuroscience, Baylor University; [b]College of Health and Human Development, Pennsylvania State University; [c]Masonic Institute for the Developing Brain, University of Minnesota; [d]Department of Psychology, Georgetown University

ABSTRACT

The compound Remote Associates Test (RAT) is a classic measure of creativity. Participants are shown three cue words (sore-shoulder-sweat) and asked to generate a word that connects them (cold). Theoretical views of RAT performance differ in the degree to which they conceptualize performance as depending on automatic spreading activation across semantic networks, strategic generation of bi-associations, and other analytical processes (e.g. executive processes that support fluid intelligence). We tested these views by adding a fourth cue word to determine whether it impaired RAT accuracy (e.g. generation of bi-associations), impaired response times (analytic processes), or improved RAT accuracy without changing response times (e.g. spreading activation). Across four experiments, 551 adults completed 3- and 4-item RAT trials that were matched on linguistic and semantic metrics. Across experiments, adding the fourth word improved accuracy by 27.91%. This performance gain occurred with either modest or no changes to response times or ratings of insight/strategy use. Interestingly, the fourth word predominantly benefited accuracy and response times on difficult trials; on easy trials, the fourth word impaired or did not change performance. The findings suggest that both automatic and strategic/analytical processes contribute to successful RAT performance, with relative dependence on these processes dynamically adapting to the demands of the individual trial.

PLAIN LANGUAGE SUMMARY

Understanding creative achievements in real-world settings requires understanding the cognitive processes that contribute to creativity. One way to do this involves examining performance on standardized creativity instruments across purposefully created conditions. We conducted four experiments in which we modified the semantic information in the Remote Associates Test (RAT), which is one of the more commonly used instruments to assess creativity. We found that as the amount of semantic information provided to participants increased, so too did performance on the RAT. There was, however, a surprising exception: on relatively easy trials, providing more semantic information sometimes hurt performance. These patterns replicated four times using different stimuli sets, designs, and participant samples. Collectively, the findings indicate that multiple cognitive processes are engaged to support creative thinking, with the relative dependence on a given process depending on the difficulty of the problem being solved. Targeting the dynamic nature of automatic and strategic/analytical thinking processes may improve the efficacy of interventions aimed at fostering creative achievements in real-world settings.

ARTICLE HISTORY
Received May 14, 2022

Introduction

Creativity is the ability to produce novel and useful ideas (for an overview on approaches to defining creativity, see Plucker, Beghetto, & Dow, 2004). Higher creative ability is associated with greater success in the classroom (Gajda, Karwowski, & Beghetto, 2017), in the workplace (Miron-Spektor, Gino, & Argote, 2011), and in overall well-being (Acar, Tadik, Myers, Van der Sman, & Uysal, 2021). Creative achievements, though, are the end-result of many cognitive processes (Green, 2016), including executive control and fluid intelligence (e.g., Frith et al., 2021; Nusbaum & Silvia, 2011). An important distinction in creative cognition research has been examining tasks that mostly require divergent thinking versus those that mostly require convergent thinking (Zhang, Sjoerds, & Hommel, 2020). Divergent thinking tasks require the generation of novel ideas to solve a problem with multiple solutions. By contrast, convergent thinking tasks require identifying the solution to a single-solution problem,

CONTACT Michael K. Scullin Michael_Scullin@Baylor.edu Department of Psychology and Neuroscience, Baylor University, One Bear Place 97334, Waco, TX 76798

This article has been corrected with minor changes. These changes do not impact the academic content of the article.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/10400419.2023.2200597.

typically those in which the solution is not immediately obvious.

The current work focuses on convergent thinking in the Remote Associates Test (RAT; Mednick, 1962). Sarnoff Mednick developed the RAT to assess his associative theory that creativity resulted from forming existing associative elements into new combinations (a view that was historically rooted in Spearman's theory of noegenesis; Fasko, 1999; Spearman, 1930). The RAT was developed to elicit domain-general convergent thinking (Ellis, Robison, & Brewer, 2021), that is, employing knowledge that, in principle, is commonly taught within society (for an alternative to the domain-general view, see Kaufman, Plucker, & Baer, 2008). Since its development, the RAT has also been used to assess insight, with respondents reporting similar amounts of subjective feelings of insight and reliance on strategies when completing RAT trials (Bowden, Jung-Beeman, Fleck, & Kounios, 2005; Chein & Weisberg, 2014).

In the original RAT, the trials' stimuli consisted of three cue words and respondents were required to find a fourth word associated with each of the three cue words either by synonymy, semantic association, or forming of a compound word. For example, in a compound RAT trial with the cues *sore-shoulder-sweat*, the best answer is *cold* (i.e., cold sore, cold shoulder, and cold sweat).

An advantage of the RAT is that each trial has a single solution, thereby ensuring ease of and uniformity in scoring. Furthermore, the administration is brief enough that it allows for multiple trials for a single respondent (Tik et al., 2018). A disadvantage of the RAT is that it may not be a so-called process-pure measure of convergent thinking (i.e. also involving some divergent thinking processes; Cortes, Weinberger, Daker, & Green, 2019); of course, the RAT is not unique in this regard as it would be difficult to support the claim that any instrument capturing a complex psychological attribute has such purity (Loevinger, 1957). In addition, some work indicates that subjective reports of insight when solving RAT problems actually reflect preparatory control rather than automatic processes (Kounios et al., 2006). Furthermore, RAT performance is sometimes poor for all but the easy trials (Bowden & Jung-Beeman, 2003b; Chein & Weisberg, 2014). This notably reduces the dispersion of RAT scores, which subsequently reduces the magnitude at which the RAT scores can predict other phenomena of interest (e.g., academic outcomes, neural correlates; Goodwin & Leech, 2006).

A goal of the current project was to determine whether RAT performance would improve – particularly on difficult trials – if a fourth cue word was added to the typical 3-item trials. In addition to having practical implications, this question is of theoretical interest because accounts of convergent thinking differ in whether they predict that a fourth word would help or harm RAT performance. For example, consider the perspective of the spreading activation model of memory in which semantic knowledge is represented as an organized network of nodes (Collins & Loftus, 1975) and creativity is determined, in part, by the connectivity across nodes (Benedek et al., 2017; Kenett, Beaty, Silvia, Anaki, & Faust, 2016; Marupaka, Iyer, & Minai, 2012). The more semantic nodes that are activated at a particular time, the more likely activation is to spread to the critical node (i.e., the solution to the problem; see Davelaar, 2015, for similar predictions from the super-additive model that emphasizes a localized semantic search space). A fourth cue word could therefore directly lead to the correct RAT solution by triggering the critical semantic node or indirectly lead to the solution by allowing incorrect solutions to be filtered (Kajić, Gosmann, Stewart, Wennekers, & Eliasmith, 2017) or otherwise dismissed (Luft, Zioga, Thompson, Banissy, & Bhattacharya, 2018; Ohlsson, 2011). Spreading activation is theorized to be an automatic process (rather than a strategic/controlled one), and consistent with this notion, success on RAT trials is often accompanied by the subjective feeling of insight ("aha!"). By the spreading activation view, adding a fourth cue word to the RAT should improve the overall performance because doing so adds additional semantic information.

Whereas the spreading activation view posits an additive account of convergent thinking, the bi-associations view posits that convergent thinking requires controlled, multiplicative processing of information. In this multiplicative view, problems are solved in two stages in which people first search for an answer and then test it against the constraints of the trial, repeating this cycle until the answer fits within all constraints (Smith, Huber, & Vul, 2013). In the context of the RAT, one must generate connections between pairs of unrelated concepts (Benedek, Jurisch, Koschutnig, Fink, & Beaty, 2020), which would involve generating each of the bi-association pairs of cue words with possible answer words, and strategically evaluating and connecting these associations to determine which answer fits all the constraints. Traditional 3-item RAT trials might be solved with this strategy, but this approach might be less advantageous in a 4-item RAT, which would require more total bi-associate iterations. Because association generation is a controlled process and (presumably) a serial process, adding a fourth cue word would introduce greater chances for errors (i.e., reduce accuracy) and slow overall responding by more than just the time required to read the additional word (~300 ms; Rayner & Clifton, 2009).

A third general view of convergent thinking is that it requires analytic, executive control processes (e.g., Lee, Huggins, & Therriault, 2014). Prior work has shown that successful RAT performance may be preceded by preparatory activation in frontal regions (Kounios et al., 2006) and language-based prior knowledge (Tik et al., 2018). Furthermore, RAT performance is often positively associated with performance on the Raven's Progressive Matrices and other intelligence instruments (e.g., Silvia, 2015; Taft & Rossiter, 1966). However, other lines of research have found the RAT to be largely explained by lexical-semantic associative processing with small or no associations to executive processes (e.g., Marko, Michalko, & Riečanský, 2019). In the current work, the analytical view would predict a fourth cue word to improve accuracy, but at the cost of slower response times (RTs).

Based on prior interpretations that the RAT is not process-pure (Cortes, Weinberger, Daker, & Green, 2019), it seems possible that respondents employ multiple processes – both relatively automatic and strategic/analytical – while solving RAT trials. By this view, adding a fourth word could sometimes help and could sometimes harm RAT performance, with the strategy/process changing as a function of the difficulty level of the trial (Valba, Gorsky, Nechaev, Tamm, & Peel, 2021). In the RAT, trial difficulty is conceptualized as the semantic distance of the solution word from the individual cue words (e.g., Marko, Michalko, & Riečanský, 2019). From this view, adding a fourth cue word to trials that have a shorter average semantic distance (easy trials) from the answer word could worsen performance because the extra semantic information would create too many options for bi-association processes. Conversely, adding a fourth cue word to trials with greater average semantic distance (difficult trials) should improve performance because the additional semantic information is necessary for identifying the critical search space in the semantic network (Davelaar, 2015).

We tested these views across four experiments that involved manipulating the addition of a fourth cue word to traditional 3-item RAT trials. The experiments were designed as incremental replications and extensions as follows: In Experiment 1, we aimed to develop an initial set of fourth cue words in which the trials would be matched to the original 3-word trials on average word length, frequency, and forward associative strength. In Experiment 2, we sought to refine the materials generated for Experiment 1 using a broad online sample, inclusive of native English and non-native English speakers. In Experiment 3, the goal was to replicate prior experiments using a sample of only native English speakers, with the methodic addition of including insight and analytic/strategy ratings following each trial. In Experiment 4, we recruited a local sample of university students and utilized a between-subjects design to examine the generalizability of effects. Experiments 1–4 were all approved by the Institutional Review Board at the corresponding author's institution and all participants provided informed consent.

Experiment 1

Methods

Participants

Participants included 110 adults (18–29 years old) who were recruited via the Prolific online research platform in April 2020 (for a similar approach, see Becker & Cabeza, 2022). An online platform was necessary due to COVID-19 remote-work requirements, but also had the benefit of allowing for samples that were not predominantly comprised of native English speakers (cf. with Experiments 3 and 4, which were solely native English speakers). Given the online platform, we screened for bot-like behavior (e.g., typing non-sense words such as *asgf*), for obvious signs of low effort (e.g., repeated RTs <1 second), and for signs that they misunderstood the task (e.g., typing cue words into the solution word box). Table S1 provides the demographic data for the 78 participants who passed the data quality screening checks.

Materials

We selected 96 compound RAT trials from the Bowden and Jung-Beeman's (2003b) norms. We generated a fourth cue word for each trial via an iterative group-consensus process (see Supplemental Appendix A). This process resulted in two lists of 96 RAT stimuli, one list consisting of the original 3-item versions and the other list containing a new, fourth word. The 3-item and 4-item versions were matched on average word length and frequency (Vitrano, Altarriba, & Leblebici-basar, 2021) as well as other key linguistic characteristics shown in Table S2 (Balota et al., 2007). In addition, the 3-item (M = .082, SD = .11) and 4-item (M = .093, SD = .11) versions were matched on forward associative strength (i.e., the average strength of the association between the cue words and the answer word; see Nelson, McEvoy, & Schreiber, 2004, norms; $t(67) = 1.429$, $p = .158$). Therefore, the new fourth item changed the quantity, not the quality (forward strength), of semantic information. Cronbach's alpha values were generally better for 4-item trials than for 3-item trials (Table S3; Taber, 2018).

Procedure

Participants first completed a bot check (captcha) to filter out potential automated responses. Then they completed demographic questionnaires and received the RAT instructions. Participants were instructed that they would view a series of words and would need to generate a new word that connects each. They were shown one example item, the solution word, and the solution word associations. Drawing on Bowden and Jung-Beeman's (2003) findings of RAT trial accuracy across multiple time limits, we set a 20-s time limit per trial. Participants were instructed to type a question mark (?) if they did not know the answer and thought they would not be able to generate it before the time limit. The question-mark response was instituted to determine how long participants would exert effort to solve the RAT trial before giving up.

During the experimental block, we manipulated the RAT item type within-subjects such that 3-item trials and 4-item trials alternated. Participants completed 30 trials (15 3-item, 15 4-item), which were randomly selected from the 96-item trial bank with the constraint that if the 3-item version appeared, then the matched 4-item version would not appear. We used the Gorilla experiment builder platform to enable RT data collection, based upon the first keypress in each trial (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020).

Statistical analysis

We used paired t-tests to compare RAT trial type (3-item, 4-item) on accuracy and RTs. RT data were separated for correctly answered trials, incorrectly answered trials, and trials that included a "?" response (26.3% of trials had no response and were excluded from RT analyses). Accuracy-RT correlations were conducted on both the participant-level (e.g., do participants with higher mean accuracy on 3-item trials also have higher mean accuracy on 4-item trials?) and the item-level (e.g., when collapsing data across all participants, are the most difficult 3-item trials also the most difficult when the fourth word is added?). Differences in degrees of freedom were due to some participants having no correctly answered trials for a trial type. Alpha was set to .05, all tests were two-tailed, and statistical analyses were conducted using SPSS version 27.

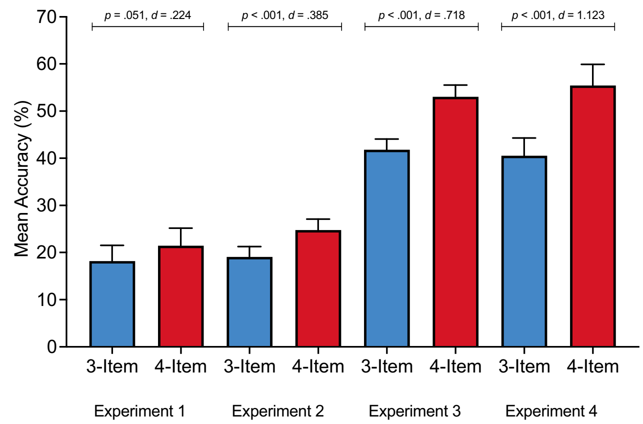

Figure 1. Mean accuracy across the 3-item and 4-item versions of the RAT in experiments 1–4. Error bars represent 95% confidence intervals.

Materials and data are available at Open Science Framework (https://osf.io/wbqzy/).

Results

Accuracy data

Accuracy on 4-item trials and 3-item trials strongly correlated in all experiments on the participant-level (see Table S4). Figure 1 shows performance on 4-item trials relative to 3-item trials, $t(77) = 1.980$, $p = .051$, $d = .224$. Inspection of individual RAT trials suggested a nominal improvement with the addition of the fourth cue word on 55.21% of trials, but a nominal worsening on 36.46% of RAT trials (8.33% showed no change; see Table S5 for these proportions across all experiments). To understand the source of this inter-item variability we examined performance relative to item-level difficulty (Valba, Gorsky, Nechaev, Tamm, & Peel, 2021). In these item-level analyses (of all 96 pairs), we estimated item-difficulty as accuracy levels on the 3-item version, and then related that original level of accuracy to the item-level difference score between the 4- and 3-item versions (higher values indicating that accuracy improved when the fourth word was added). Figure 2 illustrates that performance on the difficult RAT trials benefitted from adding a fourth word, whereas performance on the easiest trials was hindered, $r(94) = -.42$, $p < .001$. Based on Figure 1, if performance on the original 3-item RAT was below 25% accuracy, then adding a fourth word improved performance; but, if performance on the original 3-item RAT was above 40%, then adding the fourth word impaired performance.

RT data

Figure 3 shows the distribution of RTs for correctly answered trials. Most of the RAT trials were answered in 4 to 11 seconds (for a density plot, see Figure S1).

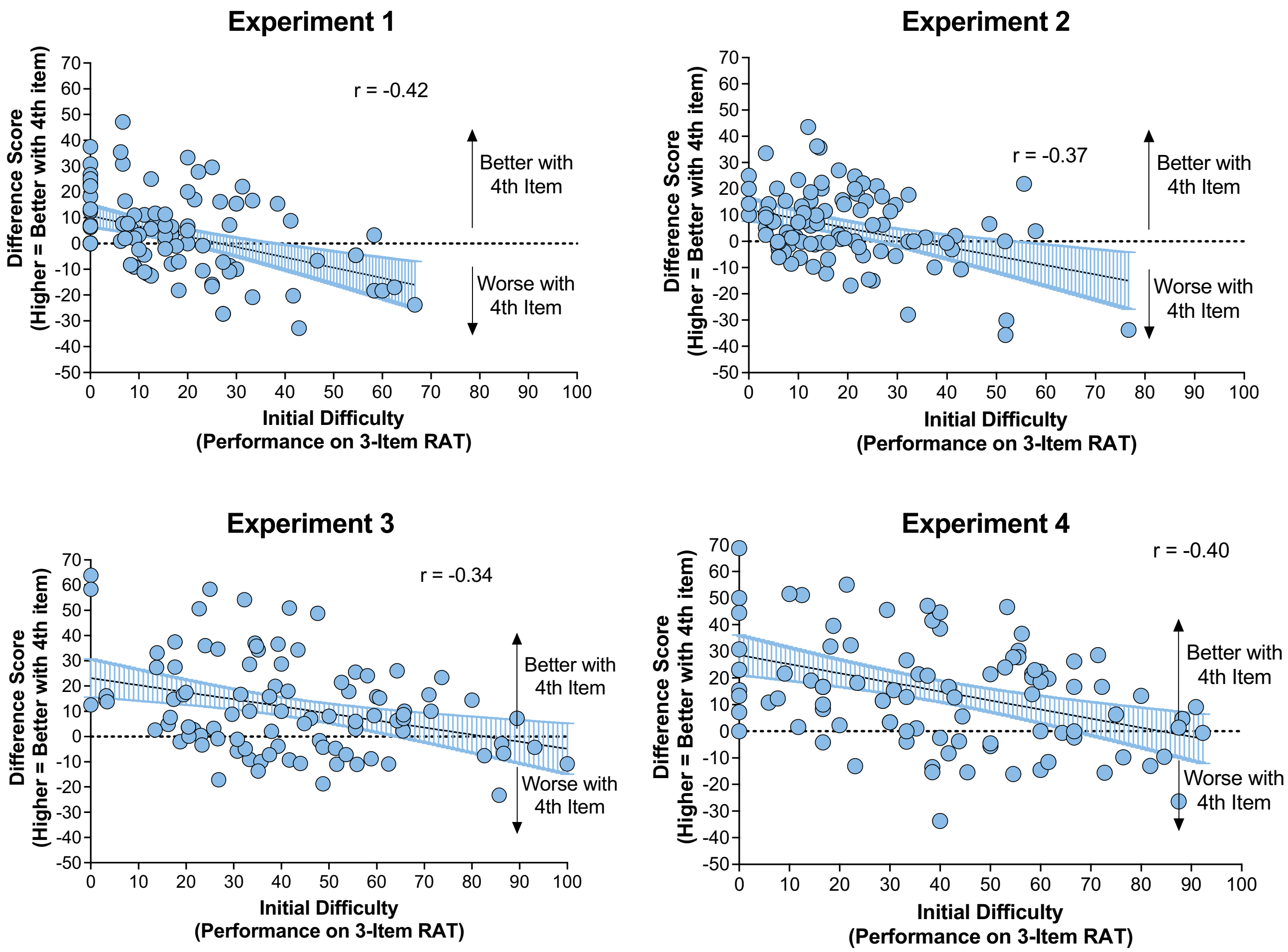

Figure 2. Scatterplots of the relationship between performance improvement when adding a fourth word (Y axis) and original difficulty level (X axis) in experiments 1–4. The dotted horizontal line indicates the level of performance that would reflect equal performance for 3-item versus 4-item versions. Upper and lower bands represent 95% confidence intervals.

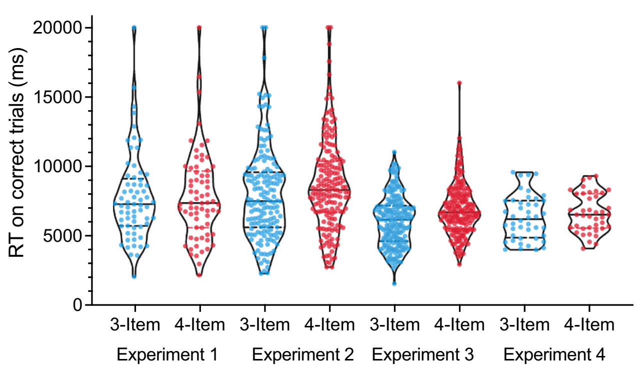

Figure 3. Violin plots of RTs on correct RAT trials in experiments 1–4. Median is marked by a solid line, quartiles are marked by dashed lines.

Participants showed similar RTs across 4-item and 3-item trials when they were correctly answered ($M_{4\text{-}item}$ = 7884 ms, SD = 3177 ms; $M_{3\text{-}item}$ = 7749 ms, SD = 3192 ms; $t(60) = .284$, $p = .778$) or incorrectly answered ($M_{4\text{-}item}$ = 9881 ms, SD = 4294 ms; $M_{3\text{-}item}$ = 9198 ms, SD = 3776 ms; $t(67) = 1.348$, $p > .182$). On trials in which participants responded with a question mark (?), RTs were significantly slower on 4-item trials (M = 11268 ms, SD = 4725 ms) than 3-item trials (M = 9994 ms, SD = 4044 ms), $t(34) = 2.040$, $p < .05$, which may indicate greater effort prior to giving up.

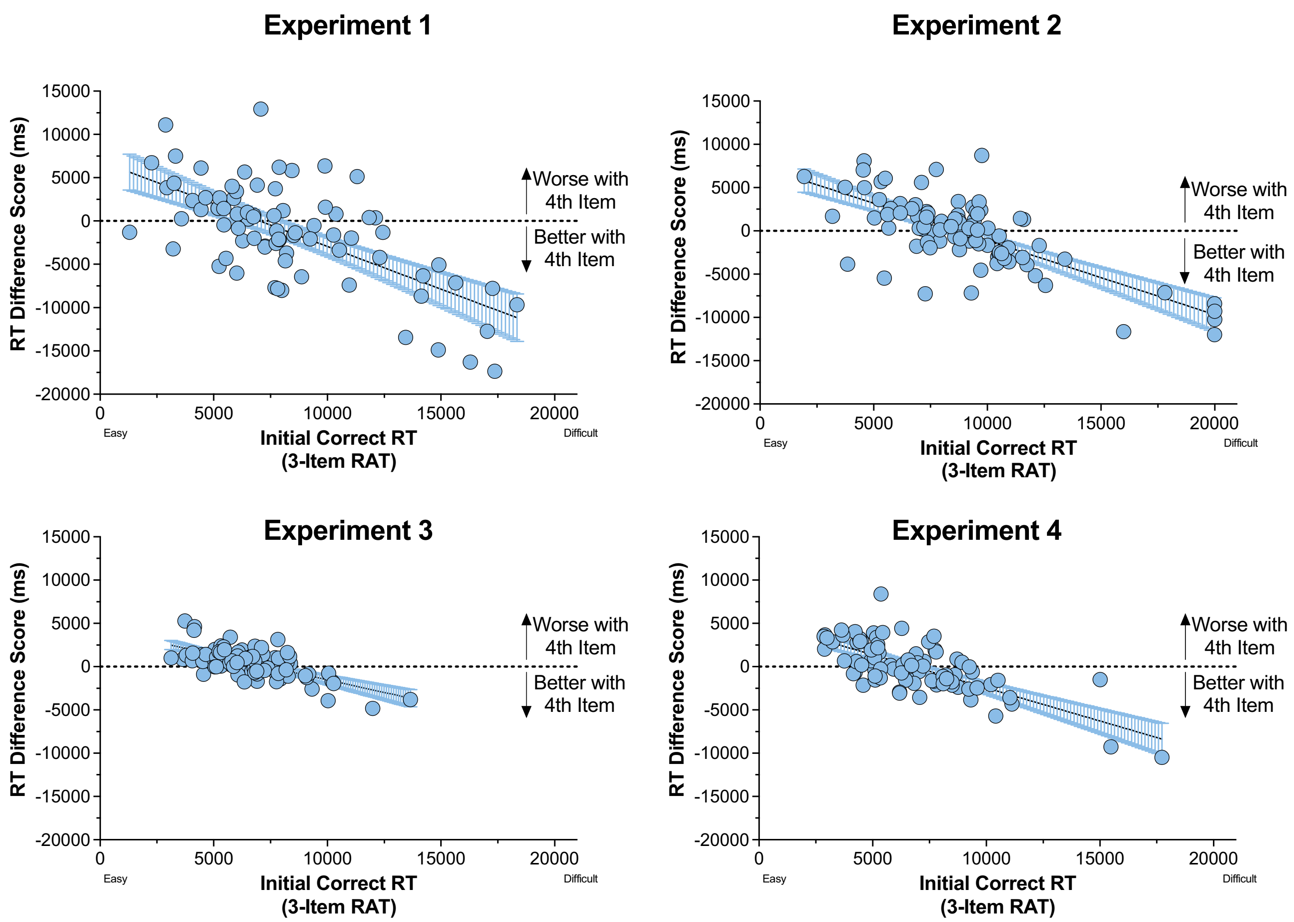

Figure 4. Scatterplots of the relationship between response time (RT) improvement when adding a fourth word (Y axis) and original difficulty level (X axis) in experiments 1–4. The dotted horizontal line indicates the level of performance that would reflect equal performance for 3-item versus 4-item versions. Upper and lower bands represent 95% confidence intervals.

On a participant-level, accuracy was not correlated with RTs (4-item: $r(68) = .039$, $p = .752$; 3-item: $r(65) = -.026$, $p = .837$), but on an item-level, greater accuracy was generally associated with faster RTs (4-item trials: $r(76) = -.271$, $p < .05$; 3-item trials: $r(75) = -.188$, $p = .102$). Consistent with our finding that item-level difficulty moderated the impact of adding a fourth item on accuracy levels, Figure 4 indicates that the longer the RT of the 3-item RAT version (which we interpret as indicating greater difficulty), the more likely adding a fourth word was to shorten RTs (represented as the item-level difference score in RT between the 4- and 3-item versions), $r(75) = -.66$, $p < .001$.

Discussion

Average performance levels on 4-item and traditional 3-item trials were relatively similar in Experiment 1 ($d = .22$). An item-level inspection though suggested that the fourth word may have benefitted difficult RAT trials while also possibly harming performance on easier RAT trials. Relative reliance on spreading activation versus strategic/analytic processes may therefore differ across trials. We aimed to replicate and extend these findings in Experiment 2.

Experiment 2

Methods

Participants

Participants were 202 adults (18–29 years old) who completed the study via the Prolific Online Research Platform in May 2020. Exclusion criteria and quality control screening were the same as in Experiment 1. Table S1 provides the demographic information for the 190 participants who passed the quality control checks.

Materials and procedure

We revised the Experiment 1 materials to replace the fourth words for 40 trials that were deemed by author consensus to be potentially misleading (Supplemental Appendix A). The goal here was to confirm that trial size (4-item versus 3-item) rather than idiosyncratic characteristics of the words selected to be the fourth items accounted for the response patterns observed in Experiment 1. The lexical characteristics are shown in Table S2. In addition, latent semantic analysis (http://wordvec.colorado.edu/) indicated no significant differences in semantic similarity of the cue words to the answer word between 3-item ($M = .208$, $SD = 0.088$) and 4-item trials ($M = .214$, $SD = 0.080$), $t(95) = -1.242$, $p = .217$.

Additional minor changes included: a) adding a second bot check (visual search task), b) adding four example problems to the RAT instructions, c) confirming all new trials were compound RAT trials, and d) adding a 500 ms interstimulus interval (fixation) to introduce brief breaks between stimuli. Statistical analyses were the same as Experiment 1.

Results

Overview

The revisions made to the procedure significantly improved the percentage of responders who passed quality control screening from 70.1% in Experiment 1 to 94.1% in Experiment 2, $X^2(1) = 31.508$, $p < .001$. Most of the patterns observed in Experiment 1 replicated in Experiment 2, typically with larger effect sizes.

Accuracy data

Figure 1 demonstrates that participants performed significantly better on 4-item trials than on 3-item trials, $t(189) = 5.301$, $p < .001$, $d = .385$. Yet, like Experiment 1, not all RAT trials showed an improvement. Adding the fourth word worsened accuracy on 35.42% of RAT trials (61.46% of trials improved and 3.12% were unchanged). Even with the updates to the trial sets, this inter-item variability was again attributable to item difficulty. Figure 2 demonstrates that the difficult RAT trials were the most likely to improve in accuracy with the addition of the fourth word, $r(94) = -.37$, $p < .001$.

RT data

For correctly answered trials, RTs on 4-item trials ($M = 8488$ ms, $SD = 3212$ ms) initially appeared slower than on 3-item trials ($M = 7986$ ms, $SD = 3336$ ms), $t(152) = 1.976$, $p = .050$, but this difference was non-significant when accounting for reading time (i.e., relative to 300 ms); $t(152) = 0.795$, $p = .428$; Rayner & Clifton (2009). Similarly, the RT difference on trials in which participants responded with a question mark (?) ($M_{4\text{-}item} = 13602$ ms, $SD = 3745$ ms; $M_{3\text{-}item} = 12867$ ms, $SD = 3957$ ms), $t(113) = 2.643$, $p < .01$, was also reduced to non-significant when accounting for reading time ($t(113) = 1.563$, $p = .121$). RTs were significantly slower on 4-item than 3-item trials that were answered incorrectly ($M_{4\text{-}item} = 12295$ ms, $SD = 4041$ ms; $M_{3\text{-}item} = 10892$ ms, $SD = 4009$ ms), $t(164) = 4.369$, $p < .001$ [even when accounting for reading time; $t(164) = 3.434$, $p < .001$], and this finding could potentially indicate a reliance on bi-associate generation. If the longer RTs reflected bi-associate generation, we would also expect RTs to positively correlate with accuracy; however, on an item-level, there was a negative correlation for 4-item trials ($r(91) = -.276$, $p < .01$) and 3-item trials ($r(90) = -.326$, $p < .01$). On a participant-level, no significant accuracy – RT associations were observed (4-item: $r(177) = -.036$, $p = .630$; 3-item: $r(160) = .190$, $p = .353$). Importantly, and mirroring the accuracy by item-level difficulty results, Figure 4 demonstrates that the longer the RT of the 3-item RAT (indicating greater difficulty), the more likely adding a fourth word was to shorten RTs, $r(90) = -.73$, $p < .001$.

Discussion

Experiment 2 showed a significant benefit of 4-item trials over traditional 3-item trials in average accuracy, with minimal changes to RTs when accounting for the additional time needed to read the fourth word. These overall patterns generally supported the spreading activation view. However, like Experiment 1, the relative reliance on automatic spreading activation versus analytical/controlled processes appeared to change dynamically across trials. In the next experiment, we extended this line of investigation to a group of only native English speakers and by querying whether correctly answered trials were being solved via subjective insight (Chein & Weisberg, 2014).

Experiment 3

Methods

Participants

Participants were 201 adults who completed the study via the Prolific platform in July 2020. To be eligible, participants needed to reside in the United States or the United Kingdom, be between the ages of 18–30, and be a native English speaker. Analyses took into account regional spelling differences (e.g. aeroplane vs. airplane). Table S1 provides the demographic information for the 199 participants who passed the quality control checks (99.0%).

Materials and procedure

Experiment 3 materials were identical to Experiment 2. The procedure was modified such that each correct response was followed by a rating of feelings of insight versus reliance on strategy when solving the trial (1 = "Complete Insight," 2 = "Partial Insight," 3 = "Partial Strategy," and 4 = "Complete Strategy;" Chein & Weisberg, 2014). Additional minor changes included: a) each trial was displayed for the entirety of the 20-s time limit to ensure that participants would not attempt to rush through trials; b) participants were shown a feedback screen that displayed the cue words, answer word, and associations; and c) to prevent fatigue, participants were given a short break after 15 trials. Statistical analyses were the same as Experiments 1–2, except that the non-parametric Wilcoxon signed-rank test was used to compare insight ratings across conditions.

Results

Accuracy data

Overall accuracy levels in this sample of only native English speakers were higher than in prior experiments and all major findings replicated, often with larger effect sizes (Figure 1). Performance was significantly better on 4-item trials than 3-item trials, $t(198) = 10.127$, $p < .001$, $d = .718$ (Figure 1). Interestingly, even with elevated performance levels in this sample, adding the fourth word was still associated with decreased accuracy on 30.21% of RAT trials (69.79% improved and 0% were unchanged). Replicating the prior experiments, RAT trials that were more difficult showed larger improvements with the fourth word, $r(94) = -.337$, $p < .001$ (Figure 2). Inspection of the 95% confidence intervals in Figure 2 indicated that the easy trials were unchanged by the fourth word (cf. Experiments 1–2 in which the fourth word worsened performance on easy trials).

RT data

Participants showed significantly slower RTs to 4-item trials when the trials were correctly answered ($M_{4\text{-}item} = 6811$ ms, $SD = 1710$ ms; $M_{3\text{-}item} = 6068$ ms, $SD = 1829$ ms; $t(194) = 4.844$, $p < .001$; reading-speed-corrected: $t(194) = 2.89$, $p = .004$). RTs for providing a question mark did not differ across trial types ($M_{4\text{-}item} = 13638$ ms, $SD = 2750$ ms; $M_{3\text{-}item} = 13407$ ms, $SD = 2271$ ms, $t(60) = 0.746$, $p = .459$) and RTs on incorrectly answered ($M_{4\text{-}item} = 11263$ ms, $SD = 3567$ ms; $M_{3\text{-}item} = 10417$ ms, $SD = 2819$ ms; $t(162) = 2.721$, $p < .01$) did not differ after accounting for reading speed ($t(162) = 1.757$, $p = .081$). On an item-level, greater accuracy was associated with faster RTs on correctly answered 4-item trials ($r(94) = -.475$, $p < .001$) and 3-item trials ($r(91) = -.498$, $p < .001$); a similar correlation was observed on a participant-level for 4-item trials ($r(197) = -.222$, $p = .002$; 3-item trials: $r(196) = -.043$, $p = .551$). Importantly, the longer the RT on the original 3-item RAT, the more likely adding a fourth word was to quicken RTs, $r(91) = -.67$, $p < .001$ (Figure 4).

Insight ratings data

Figure 5a illustrates that participants predominantly reported "Partial Insight" or "Complete Insight" on 3-item trials (70.35%) and 4-item trials (70.08%). A Wilcoxon signed-ranks test showed no significant differences in the mean ratings between trial types ($M_{3\text{-}item} = 1.97$, $SD = .62$, $M_{4\text{-}item} = 1.94$, $SD = .59$; $Z = 0.542$, $p = .59$). On an item-level, trials with greater accuracy levels were also those more likely to be solved with insight, both in the 4-item sets ($r(94) = -.22$, $p < .05$) and the 3-item sets ($r(91) = -.50$, $p < .001$). On a participant-level, the accuracy – insight associations were negligible (4-item: $r(197) = .086$, $p = .231$; 3-item: $r(196) = .077$, $p = .286$).

Discussion

Adding a fourth word improved RAT performance, particularly for the more difficult RAT trials. Furthermore, participants reported solving RAT problems with subjective insight more often than relying on strategies (note, however, that process-based interpretations of subjective insight remain debated and could actually reflect preparatory executive control processes: Becker, Wiedemann, & Kühn, 2020; Bowden & Jung-Beeman, 2003a; Chein & Weisberg, 2014; Jung-Beeman et al., 2004; Kounios et al., 2006). As a final replication and extension, for Experiment 4, we recruited local university students and manipulated trial-type between-subjects rather than within-subjects. Using a between-subjects design is important for identifying which, if any, of the patterns in Experiments 1–3 could be attributable to task switching between 3-item and 4-item trial types.

Experiment 4

Methods

Participants

Eighty-five undergraduate students were recruited via SONA in fall 2020 and received partial class credit for

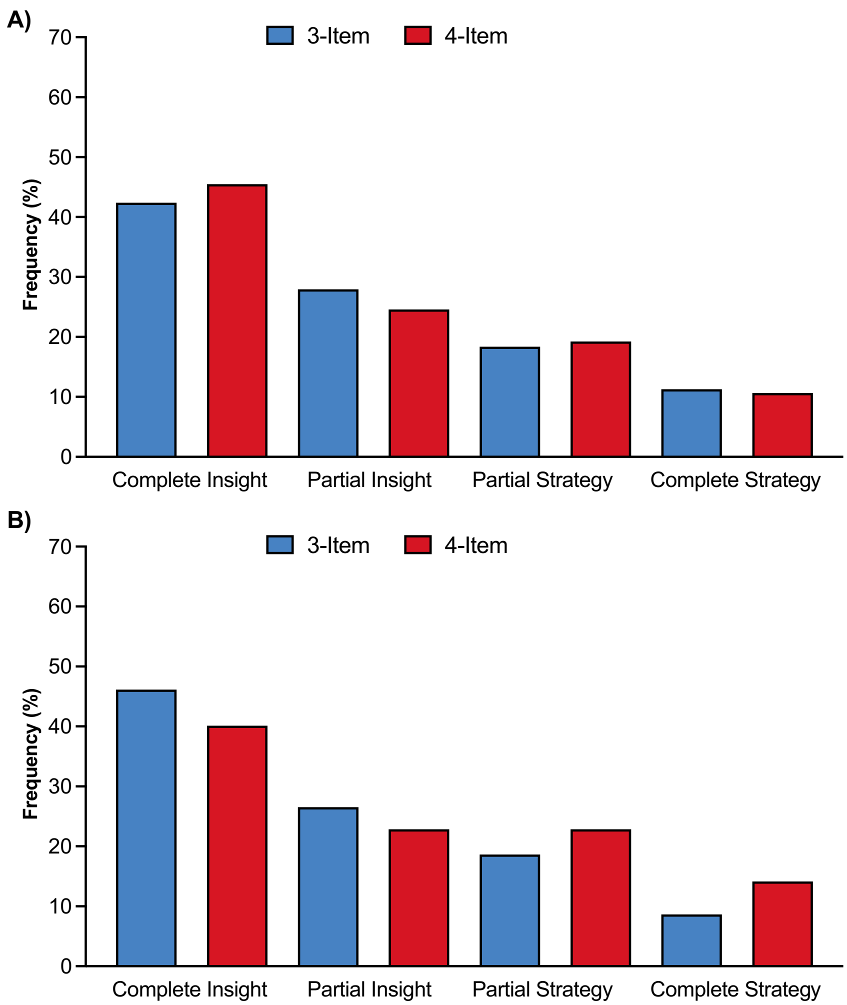

Figure 5. Frequency of insight rating responses in experiment 3 (A) and experiment 4 (B).

participating. The sample size was determined by an a priori power analysis that was based on the effect size for the primary dependent variable in Experiment 3 ($d = .718$), a two-tailed test, and alpha set to 0.05. This test indicated that 84 participants were needed to achieve .90 power to detect condition effects on accuracy. All participants (100%) passed quality control screening (Table S1).

Procedure

All materials and procedures were the same as Experiment 3 except that participants were randomly assigned to solve only 3-item trials or only 4-item trials (between-subjects design). The experimental block was 30 trials in length.

Results

Accuracy data

Figure 1 demonstrates that participants in the 4-item condition performed significantly better than participants in the 3-item condition, $t(83) = 5.175$, $p < .001$, and with an effect size exceeding each of the prior experiments ($d = 1.123$). Most RAT trials improved in mean accuracy with the fourth word added (71.88%), though some still worsened in accuracy (23.96%; 4.16% were unchanged). RAT trials that were more difficult in their 3-item version were more likely to improve in accuracy when adding a fourth word, $r(94) = -.402$, $p < .001$ (Figure 2).

RT data

RTs were generally quicker in this university student sample (relative to Experiments 1–3). There were no condition differences in RTs for correctly answered trials ($M_{4\text{-}item} = 6646$ ms, $SD = 1388$ ms; $M_{3\text{-}item} = 6323$ ms, $SD = 1608$ ms; $t(83) = 0.987$, $p = .327$), incorrectly answered trials ($M_{4\text{-}item} = 11173$ ms, $SD = 2753$ ms; $M_{3\text{-}item} = 10189$ ms, $SD = 2421$ ms; $t(82) = 1.742$, $p = .085$), or trials that included a question mark (?) response ($M_{4\text{-}item} = 13361$ ms, $SD = 1478$ ms; $M_{3\text{-}item} = 13822$ ms, $SD = 3855$ ms; $t(24) =$

−0.360, $p = .722$). Accuracy – RT associations were negligible on a participant-level (4-item: $r(41) = -.119$, $p = .457$; 3-item: $r(44) = -.108$, $p = .486$). On an item-level, accuracy and RT showed a significant inverse correlation for 3-item trials ($r(84) = -.376$, $p < .001$) and 4-item trials ($r(93) = -.314$, $p < .01$), consistent with Experiments 1–3. In addition, adding a fourth word was most likely to quicken RTs on the most difficult 3-item RAT trials (i.e., those with the slowest RTs), $r(84) = -.72$, $p < .001$ (Figure 4).

Insight data

Figure 5b shows that there were no significant differences in mean insight ratings between 3-item trials ($M = 1.89$, $SD = .50$) and 4-item trials ($M = 2.10$, $SD = .61$; $Z = 1.805$, $p = .071$). Greater ratings of insight occurred on trials with higher levels of accuracy on an item-level in the 4-item condition ($r(93) = -.45$, $p < .001$; 3-item condition: $r(84) = -.19$, $p = .084$). This association was not observed at the participant-level (4-item: $r(41) = .080$, $p = .618$; 3-item: $r(44) = .040$, $p = .799$).

Discussion

Experiment 4 replicated the major findings with a between-subjects design: adding a fourth item improved RAT performance, especially for the most difficult trials. Experiment 4 showed that these benefits could be achieved without changing the nature of the task (insight ratings) or incurring RT costs. One potential explanation for RT slowing across some conditions in Experiments 2–3 (though not always exceeding the additional 300 ms needed to read the fourth word) was that alternating trial type (within subjects) led to task-switching costs for the trial type that required the most reading (4-item trials). When comparing the outcomes of Experiment 4 to the preceding experiments, it appears that the benefits of a 4-item RAT for increasing spreading activation across relevant semantic nodes is greatest amongst native English speakers (highly connected semantic networks) and when using a between-subjects design (eliminating task switching costs).

General discussion

Given the difficulty participants have in identifying connections between three cue words when there are time constraints (Bowden & Jung-Beeman, 2003b), one might reasonably expect that adding a fourth word would only make the RAT more difficult. Yet, participants consistently performed better on 4-item RAT trials than on the traditional 3-item RAT trials, by an average percent increase of 27.91%. The benefit of the fourth item replicated in native and non-native English speakers (Experiments 1–2 vs. 3–4), within- and between-subjects designs (Experiments 1–3 vs. 4), broad online samples and specific university samples (Experiments 1–3 vs. 4), and multiple sets of materials (i.e. iterative changes between each experiment), with the largest effects occurring for native English speakers in between-subjects designs (Experiment 4). Additionally, the fourth word was found to differentially benefit trial performance based on trial difficulty across all experiments. In this section, we consider the processes by which the fourth word influenced convergent thinking.

Semantic processes contributing to RAT performance

According to the spreading activation view, semantic knowledge is organized within a network of nodes (Collins & Loftus, 1975), and lexical processing automatically triggers the activation that spreads across associated nodes (Friedrich, Henik, & Tzelgov, 1991). One interpretation of the current findings, therefore, was that the additional lexical/semantic information (fourth word) increased the likelihood of activation spreading to the critical node, resulting in retrieval of the RAT solution. Reinforcing this interpretation, the effect sizes for adding the fourth word *tripled* from Experiments 1–2 in which non-native English speakers were eligible to Experiments 3–4 in which only native English speakers were recruited (i.e., individuals who have stronger semantic network connectivity for the verbal materials).

Other interpretations are also possible, such as the fourth word leading participants to consider and then inhibit incorrect solutions (White & Shah, 2006) or engage other executive control processes (Lee & Therriault, 2013). These executive processes could be operating even in cases in which participants report solving trials with self-reported “aha” insight (Kounios et al., 2006). However, by this analytical/executive view, we would expect to see RT slowing on correct trials to a level that exceeds the time necessary to read the extra word; such slowing was usually not observed, and there was specifically no evidence for slowing in Experiment 4, which used a between-subjects manipulation (thereby eliminating potential task switching confounds).

Perhaps more likely was that adding a fourth word helped to constrain the semantic neighborhood, thereby facilitating automatic spreading activation processes or increasing the efficiency of a search process within a localized space (Kounios et al., 2006). This view converges with Davelaar’s (2015) superadditive model that a search process can be localized to the intersection of the RAT cue words and to nodes that have strong similarities to the cue words. Interestingly, Davelaar's model theorizes this process to be most effective with weak targets, which was consistent with our finding that the fourth word was most beneficial on relatively difficult RAT trials. However, it is important to acknowledge that in our work, a variable/dimension that was not built into the stimulus set was the likelihood of search intersection. As an example, "sore" and "sweat" may lead to a direct intersection search for "cold" whereas "shoulder" may not lead to such a direct search intersection. Therefore, the RAT may be an imperfect measure of spreading activation because some stimuli may have a greater likelihood of search intersection than others.

Dynamic processes during convergent thinking

Though average RAT performance consistently improved with a fourth word, some individual trials did not. Performance actually worsened for 23% to 36% of the RAT problems, even though multiple sets of materials were used across experiments. This finding was consistent with the prediction of the bi-associations view that convergent thinking can involve controlled, multiplicative processing of information (Smith, Huber, & Vul, 2013) and less consistent with analytical processing views that assume performing the RAT is congruent with solving fluid intelligence tasks (e.g., see Lee, Huggins, & Therriault, 2014). We observed that performance was most likely to worsen when the fourth word was added to relatively easy RAT trials. On such trials, the first cue word often has a short semantic distance from the solution and introducing a fourth word would increase the total iterations required for checking the bi-association pairs (resulting in more opportunities for cognitive errors). Because the fourth word benefitted performance on difficult trials and hurt performance on easy trials, it appears that the relative reliance on automatic spreading activation processes and controlled processes can change dynamically within a single task block.

Valba, Gorsky, Nechaev, Tamm, and Peel (2021) similarly proposed that RAT strategy changes across relatively easy and relatively difficult trials. Specifically, they argued that easy trials are best solved by removing all weak associations and only identifying strong ones; applied to our current findings, performance on relatively easy RAT trials may have worsened with the fourth word because it was yet another weak association (or at best, a redundant strong association). In contrast to the pattern for easy trials, Valba et al. argued that difficult trials are best solved via continued processing of moderately weak associations. According to this viewpoint, a fourth word would lead to the correct RAT solution by providing an additional source of moderately weak information. Thus, like other areas in the cognitive sciences, the strategy used to perform creativity tasks can fluctuate on a trial by trial basis (e.g., Shelton & Scullin, 2017).

Limitations and conclusions

For 60 years, the RAT test has been a useful measure of convergent thinking (Mednick, 1962). Its advantages have included being objective, domain-general, and brief. Its disadvantages have included reliance on self-reported insight/strategy use, language dependency (cf. Becker & Cabeza, 2022), and showing low levels of performance on difficult trials. Because difficult trials may involve qualitatively different cognitive processes than easy trials (e.g., Valba, Gorsky, Nechaev, Tamm, & Peel, 2021), and because floor performance on such trials limits the efficacy of studying the underlying mechanisms and real-world correlates (Goodwin & Leech, 2006), there is a need to shift-the-scale of performance. The current experiments provide replicable evidence that adding a fourth word to difficult RAT trials improves overall performance, without changing the nature of the task (self-reported insight and RTs). In addition, the current work indicates that multiple cognitive processes contribute to successful convergent thinking, with the relative reliance on relatively automatic versus strategic/controlled processes changing dynamically with the demands of the task.

Disclosure statement

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

Funding

This research was supported by the National Science Foundation under grant #1920730.

ORCID

Michael K. Scullin http://orcid.org/0000-0002-7578-7587

Data availability statement

All data are publicly available at https://osf.io/wbqzy/

References

Acar, S., Tadik, H., Myers, D., Van der Sman, C., & Uysal, R. (2021). Creativity and well-being: A meta-analysis. *The Journal of Creative Behavior*, *55*(3), 738–751. doi:10.1002/jocb.485

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. doi:10.3758/s13428-019-01237-x

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B. . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. doi:10.3758/bf03193014

Becker, M., & Cabeza, R. (2022). *Assessing creativity independently of language: A language-independent remote associate task (LI-RAT)*. Behavior Research Methods. doi:10.3758/s13428-021-01773-5

Becker, M., Wiedemann, G., & Kühn, S. (2020). Quantifying insightful problem solving: A modified compound remote associates paradigm using lexical priming to parametrically modulate different sources of task difficulty. *Psychological Research*, *84*(2), 528–545. doi:10.1007/s00426-018-1042-3

Benedek, M., Jurisch, J., Koschutnig, K., Fink, A., & Beaty, R. E. (2020). Elements of creative thought: Investigating the cognitive and neural correlates of association and bi-association processes. *Neuroimage*, *210*, 116586. doi:10.1016/j.neuroimage.2020.116586

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

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

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

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, *9*(7), 322–328. doi:10.1016/j.tics.2005.05.012

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. doi: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. doi:10.1037/0033-295X.82.6.407

Cortes, R. A., Weinberger, A. B., Daker, R. J., & Green, A. E. (2019). Re-examining prominent measures of divergent and convergent creativity. *Current Opinion in Behavioral Sciences*, *27*, 90–93. doi:10.1016/j.cobeha.2018.09.017

Davelaar, E. J. (2015). Semantic search in the remote associates test. *Topics in Cognitive Science*, *7*(3), 494–512. doi:10.1111/tops.12146

Ellis, D. M., Robison, M. K., & Brewer, G. A. (2021). The cognitive underpinnings of multiply-constrained problem solving. *Journal of Intelligence*, *9*(1), 7. doi:10.3390/jintelligence9010007

Fasko, D. (1999). Associative theory. *Encyclopedia of Creativity*, *1*, 135–139.

Friedrich, F. J., Henik, A., & Tzelgov, J. (1991). Automatic processes in lexical access and spreading activation. *Journal of Experimental Psychology Human Perception and Performance*, *17*(3), 792–806. doi:10.1037/0096-1523.17.3.792

Frith, E., Kane, M. J., Welhaf, M. S., Christensen, A. P., Silvia, P. J., & Beaty, R. E. (2021). Keeping creativity under control: Contributions of attention control and fluid intelligence to divergent thinking. *Creativity Research Journal*, *33*(2), 138–157. doi:10.1080/10400419.2020.1855906

Gajda, A., Karwowski, M., & Beghetto, R. A. (2017). Creativity and academic achievement: A meta-analysis. *Journal of Educational Psychology*, *109*(2), 269–299. doi:10.1037/edu0000133

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *Journal of Experimental Education*, *74*(3), 251–266. doi:10.3200/JEXE.74.3.249-266

Green, A. E. (2016). Creativity, within reason: Semantic distance and dynamic state creativity in relational thinking and reasoning. *Current Directions in Psychological Science*, *25*(1), 28–35. doi:10.1177/0963721415618485

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. doi:10.1371/JOURNAL.PBIO.0020097

Kajić, I., Gosmann, J., Stewart, T. C., Wennekers, T., & Eliasmith, C. (2017). A spiking neuron model of word associations for the remote associates test. *Frontiers in Psychology*, *8*, 99. doi:10.3389/fpsyg.2017.00099

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. Hoboken, NJ: John Wiley & Sons.

Kenett, Y., Beaty, R., Silvia, P., Anaki, D., & Faust, M. (2016). Structure and flexibility: Investigating the relation between the structure of the mental lexicon, fluid intelligence, and creative achievement. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(4), 377–388. doi:10.1037/aca0000056

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. doi:10.1111/j.1467-9280.2006.01798.x

Lee, C. S., Huggins, A. C., & Therriault, D. J. (2014). A measure of creativity or intelligence? Examining internal and external structure validity evidence of the remote associates test. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(4), 446–460. doi:10.1037/a0036773

Lee, C. S., & Therriault, D. J. (2013). The cognitive underpinnings of creative thought: A latent variable analysis exploring the roles of intelligence and working memory in three creative thinking processes. *Intelligence*, *41*(5), 306–320. doi:10.1016/j.intell.2013.04.008

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. doi:10.2466/pr0.1957.3.3.635

Luft, C. D. B., Zioga, I., Thompson, N. M., Banissy, M. J., & Bhattacharya, J. (2018). Right temporal alpha oscillations as a neural mechanism for inhibiting obvious associations. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(52), E12144–12152.

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

Marupaka, N., Iyer, L. R., & Minai, A. A. (2012). Connectivity and thought: The influence of semantic network structure in a neurodynamical model of thinking. *Neural Networks*, *32*, 147–158. doi:10.1016/j.neunet.2012.02.004

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

Miron-Spektor, E., Gino, F., & Argote, L. (2011). Paradoxical frames and creative sparks: Enhancing individual creativity through conflict and integration. *Organizational Behavior and Human Decision Processes*, *116*(2), 229–240. doi:10.1016/j.obhdp.2011.03.006

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. doi:10.3758/BF03195588

Nusbaum, E. C., & Silvia, P. J. (2011). Are intelligence and creativity really so different?: Fluid intelligence, executive processes, and strategy use in divergent thinking. *Intelligence*, *39*(1), 36–45. doi:10.1016/j.intell.2010.11.002

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. New York, NY: Cambridge University Press.

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist*, *39*(2), 83–96. doi:10.1207/s15326985ep3902_1

Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, *80*(1), 4–9. doi:10.1016/j.biopsycho.2008.05.002

Shelton, J. T., & Scullin, M. K. (2017). The dynamic interplay between bottom-up and top-down processes supporting prospective remembering. *Current Directions in Psychological Science*, *26*(4), 352–358. doi:10.1177/0963721417700504

Silvia, P. J. (2015). Intelligence and creativity are pretty similar after all. *Educational Psychology Review*, *27*(4), 599–606. doi:10.1007/s10648-015-9299-1

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the remote associates test. *Cognition*, *128*(1), 64–75. doi:10.1016/j.cognition.2013.03.001

Spearman, C. E. (1930). *Creative mind*. London, England: Cambridge University Press.

Taber, K. S. (2018). The use of cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. doi:10.1007/s11165-016-9602-2

Taft, R., & Rossiter, J. R. (1966). The remote associates test: Divergent or convergent thinking? *Psychological Reports*, *19*, 1313–1314. doi:10.2466/pr0.1966.19.3f.1313

Tik, M., Sladky, R., Luft, C. D. B., Willinger, D., Hoffmann, A., Banissy, M. J. . . . Windischberger, C. (2018). Ultra-high-field fMRI insights on insight: Neural correlates of the aha!-moment. *Human Brain Mapping*, *39*(8), 3241–3252. doi:10.1002/hbm.24073

Valba, O., Gorsky, A., Nechaev, S., Tamm, M., & Peel, L. (2021). Analysis of English free association network reveals mechanisms of efficient solution of remote association tests. *PLoS One*, *16*(4), e0248986. doi:10.1371/journal.pone.0248986

Vitrano, D., Altarriba, J., & Leblebici-basar, D. (2021). Revisiting mednick's (1962) theory of creativity with a composite measure of creativity: The effect of stimulus type on word association production. *The Journal of Creative Behavior*, *55*(4), 925–936. doi:10.1002/jocb.498

White, H. A., & Shah, P. (2006). Uninhibited imaginations: Creativity in adults with attention-deficit/hyperactivity disorder. *Personality and Individual Differences*, *40*(6), 1121–1131. doi:10.1016/j.paid.2005.11.007

Zhang, W., Sjoerds, Z., & Hommel, B. (2020). Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. *NeuroImage*, *210*, 116572. doi:10.1016/j.neuroimage.2020.116572