JOURNAL OF
Creative Behavior
Publication of the Creative Education Foundation™

MARIE BEISEMANN
BORIS FORTHMANN
PAUL-CHRISTIAN BÜRKNER
HEINZ HOLLING

Psychometric Evaluation of an Alternate Scoring for the Remote Associates Test

ABSTRACT

The Remote Associates Test (RAT; Mednick, 1962; Mednick & Mednick, 1967) is a commonly employed test of creative convergent thinking. The RAT is scored with a dichotomous scoring, scoring correct answers as 1 and all other answers as 0. Based on recent research into the information processing underlying RAT performance, we argued that the dichotomous scoring may lead to a loss of potentially relevant information. Thus, we proposed an alternate scoring based on semantic similarity between the answer given by the participant and the correct solution using Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). We evaluate the psychometric properties of the alternate LSA scoring and found evidence of construct validity for the LSA scoring which was comparable to findings for the standard scoring, but not better as we would have expected. Thus, our expectations that LSA-based scoring of the RAT counteracts potential information loss were not met. However, LSA based scorings appear to be a promising alternative for hardly solvable RAT items. We conducted additional analyses comparing different RAT item types with regard to their validity as well as evaluating the information uniquely contained in the LSA scoring. Implications of all finding for existing research using RAT items are discussed.

Keywords: Remote Associates Test, Latent Semantic Analysis, compound RAT, convergent thinking, creative thinking.

The Remote Associates Test (RAT; Mednick, 1962; Mednick & Mednick, 1967) is one of the most commonly used tests to measure creative thinking (Kaufman, Plucker & Baer, 2008). For example, the RAT has been translated into several other languages, such as French (Chun & Hupé, 2016), Turkish (Özen, Doğan & Cinan, 2015), and Italian (Salvi, Costantini, Bricolo, Perugini & Beeman, 2016). Each item consists of three seemingly unrelated cue words; the solution to each item is the word that connects all three cue words. Consider for example the item (Mednick, 1962): rat—blue—cottage; the solution to this item would be cheese as it is associated with each one of the cue words and thus provides a connecting link between them. Based on research investigating the information processes underlying RAT performance (Smith, Huber & Vul, 2013; Smith & Vul, 2015), we evaluated an alternative scoring for the RAT which quantifies how semantically similar an answer was as compared to the designated correct solution and thereby offers potentially valuable information about inter-individual differences in RAT performance.

THE REMOTE ASSOCIATES TEST

On the basis of prominent case studies (e.g., Albert Einstein), Mednick (1962) defined creative thinking as “the forming of associative elements into new combinations which either meet specified requirements or are in some way useful” (p. 221). Based on this idea, he developed a test for inter-individual differences in creative thinking: the RAT (Mednick & Mednick, 1967). Mednick originally developed two college-level versions of the RAT with a forty-minute time limit. As mentioned above, the test requires subjects to find an associative connection between three seemingly unrelated stimuli (i.e., words).

Over the years, the RAT has been used to examine a variety of research questions from several different fields of research. Originally, Mednick developed the RAT as an operationalization of his associative theory of creative thinking and employed the test as such, for example, examining the effects of specific associative priming on RAT performance (Mednick, Mednick & Mednick, 1964). Later on, social psychologists used the

RAT to investigate social decision processes (Laughlin, Kerr, Munch & Haggarty, 1976). RAT performance —as measured using Mednick and Mednick's (1967) original items or items which resemble the original items in structure and instruction—has been linked to insight problem-solving (*r* = .37 in Schooler & Melcher, 1995; *r* = .42 in Ansburg, 2000). After developing and providing normative data for 144 compound RAT items (CRAT; Bowden & Jung-Beeman, 2003), those CRAT items were used to investigate the neural processes and activation during the experience of insight in problem-solving (Bowden, Jung-Beeman, Fleck & Kounios, 2005). However, CRAT items differ slightly from Mednick's (1962) original associational RAT (ARAT) items: the solution word is associated with all three item words in the same manner—formation of a compound word.

Researchers studying intuition have also used a slightly altered version of the RAT (Balas, Sweklej, Pochwatko & Godlewska, 2012; Bolte & Goschke, 2005). Their research is concerned with coherence judgments based on intuition and factors influencing those judgments. They had participants answer RAT items, half of them actual ARAT items (coherent trials) and the other half word triplets that look just like ARAT items but there is no correct solution which is associated with all three cue words (incoherent trials). Participants had to attempt to solve items under time limitations and judge whether each item was coherent or not. As they were interested in intuition, the researchers excluded solved items from their analyses. Diverging from how the RAT is typically scored, they deemed ARAT items solved when they were answered with the correct answer, a synonym for the correct answer, or an unanticipated answer which independent raters considered semantically closely related to the three clue words. This approach suggests that there are differences in answers not identical to the designated solution, in particular it suggests that semantically similar answers to the solution are "more correct" than other answers.

Most recently, the information processing underlying RAT performance has been studied. Different computational models of the RAT have been developed whose performances predict human performance on the RAT (Gupta, Jang, Mednick & Huber, 2012; Kajić, Gosmann, Stewart, Wennekers & Eliasmith, 2017; Olteţeanu & Falomir, 2015). These computational models help to identify factors which determine item difficulty for human participants. Research has singled out word frequency (as a proxy for associative strength) as an important factor (Gupta et al., 2012; Olteţeanu & Falomir, 2015), and the ratio between the desired and the total number of associations (Oltețeanu & Schultheis, 2017). Another approach to investigating the information processes underlying RAT performance was taken by having human participants document each word they considered as an answer to the RAT items (Davelaar, 2015; Smith & Vul, 2015; Smith et al., 2013). Smith and Vul (2015) related considered answers, cue words, and solutions to each other using Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) which provides measures of semantic similarity between words. They found that while searching for the solution to a RAT item, participants consider and discard different answers which—as long as they are on the right path toward the correct solution—become increasingly more semantically similar to the designated solution. In contrast with other accounts, the authors interpret exhibited search patterns as evidence of a sequential search strategy: Which word will be considered next as the solution to a RAT item depends on the previously considered word, with the considered answers being semantically related to each other and ideally to the designated solution. Furthermore, Akbari Chermahini, Hickendorff and Hommel (2012) found that RAT scores were significantly related to insight problems (*r* = .39) and Raven's Advanced Progressive Matrices (*r* = .47) but not to divergent thinking. Understanding the RAT as a measure of convergent thinking, they concluded from the former that there was evidence for the RAT's convergent validity and from the latter that there was evidence for the RAT's discriminative validity. Lee, Huggins and Therriault (2014) found RAT scores to be related to insight problems (*r* = .14–.18), working memory (*r* = .17–.25), processing speed (*r* = .17), fluid intelligence (*r* = .34), a vocabulary task (*r* = .42), and grade point average (*r* = .16), concluding that their results provide sufficient evidence for the RAT's convergent validity. In addition, they found RAT performance to be unrelated to either measures of divergent thinking or the personality trait openness to experience.

THE PRESENT RESEARCH

In most studies, the RAT is scored with a dichotomous scoring: a 1 for correct and a 0 for incorrect answers. Summing across all items yields the total RAT score for a person. Another way in which the RAT has been scored stems from research investigating intuition (Balas et al., 2012; Bolte & Goschke, 2005): RAT items were also considered solved when answered with a synonym for the correct solution or a solution which was rated semantically related to the three cue words. Even though the intent behind this scoring method was to be as conservative as possible when it came to the decision of what was considered an unsolved item (Bolte & Goschke, 2005), it does suggest that there are answers to RAT items which do not correspond exactly to the designated solution but cannot really be considered entirely wrong either. Another implication is that ratings of semantic association between participants' answers and the three cue words have been deemed appropriate means to identify such answers.

Previous research into the information processing underlying RAT performance has indicated that when searching for a solution for a RAT item, people employ a local search strategy with sequential dependence (Smith et al., 2013). They run through a sequence of guesses at the correct answer, considering each guess and either discard it and continue considering new guesses, or use the considered guess as their response. Sequential dependence means that in this process, they base their newest guess at the solution on their previous guess. This sequential dependence was found for solved and unsolved items which suggests a random walk in semantic space. Importantly, Smith et al. (2013) excluded unsolved items from analyses in which participants had marked a wrong answer which they believed to be the solution. This case, however, is crucial for the current work. We hypothesize individual differences in semantic similarity between the solution and final responses (no matter if correct or incorrect). These individual differences could emerge (besides other possible mechanisms) when choosing starting points for random walks (e.g., choosing starting points with a greater likelihood to come close to the solution).

Furthermore, the RAT is often administered under time limits, such as one minute. After lapse of that minute, the RAT does not offer any information about whether a person was generally unable to generate the correct answer or unable to generate the correct answer in time. In our opinion, that constitutes a loss of potentially relevant information for differentiating between participants. We argue that by only scoring the RAT for correct and incorrect answers, an artificial dichotomization is created. We propose an alternate scoring which would account for such information which is potentially lost on the standard scoring. By measuring the semantic similarity of a person's response to the correct answer, we would still differentiate between those who answered correctly and those who answered incorrectly (i.e., moving into an incorrect direction). However, we would have a measure that also credits having taken the right approach to the correct answer (but failing to reach it in time), offering additional information in comparison to the standard RAT scoring.

In order to measure the semantic similarity of a response to the correct answer, we employed Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). LSA has been previously used in relation to the RAT in order to examine semantic similarity between item words, solution words, and/or considered potential solutions in the context of examining the information processing underlying RAT performance (e.g., Davelaar, 2015; Gupta et al., 2012; Olteţeanu & Falomir, 2015; Smith et al., 2013). In short, LSA provides a high-dimensional semantic space which expresses relations between words or concepts and contexts (i.e., documents) in which they appear more or less commonly (Landauer & Dumais, 1997). For a more elaborate yet simple introduction to LSA as a method applied in creativity research, please consult Forthmann, Oyebade, Ojo, Günther and Holling (2018). The semantic similarity between two words in the semantic space can be expressed as the cosine between the vectors for those two words in the semantic space (Landauer & Dumais, 1997). Essentially, two words are then semantically more similar, if they commonly appear in the same contexts. Depending on the methods used within LSA to obtain the semantic space, the cosine similarity between any two words either ranges from -1 to 1 or from 0 to 1 (Günther & Marelli, 2016), 1 indicating in both cases that the two words are the same. The greater the cosine value between two words, the more semantically related they are. With our proposed LSA scoring, the cosine similarity between the participant's response and the correct solution constitutes that person's score for that item.

The aim of the present study was to use LSA as an alternative scoring for the RAT. To that purpose and following previous psychometric research (Akbari Chermahini et al., 2012; Lee et al., 2014), we compared the LSA scoring's psychometric properties to the standard scoring's psychometric properties. Validity was assessed in the same manner as in previous research, relating both LSA and standard RAT scores to convergent thinking tasks (e.g., performance on insight problems, measures of intelligence) as well as contrasting them with measures of divergent thinking. Previous studies (Akbari Chermahini et al., 2012; Lee et al., 2014) only used performance on insight problems as an indicator for problem-solving ability. In order to be able to corroborate any findings regarding this construct within the same study, we chose to additionally use an alternative measure (i.e., a self-report) to assess problem-solving ability. Using the self-report measure we chose (Productive-Reproductive Thinking Inventory; Cunningham & MacGregor, 2016) in a validation study for the RAT was also interesting because it had not previously been done. We additionally related both types of RAT scores to measures of (current) achievement motivation, expecting—as has been shown for other cognitive tasks (Freund, Kuhn & Holling, 2011)—both RAT scores to be positively related to the interest facet of current achievement motivation.

METHOD

PARTICIPANTS

Our sample consisted of 202 participants of whom we excluded 87 incomplete test-sessions as well as two participants who did not give us consent to analyze their data and two participants whose native language was not German. One participant appeared to have partaken in our study twice. We excluded the respective participant's second trial. The remaining 111 participants (89 females, 21 males, one not specified) were included in our analyses. Participants' age ranged from 18 to 55 years (M = 22.92; SD = 5.84). Our sample included mostly university students (94.6%) of which 95.2% studied Psychology. This high level of education in our sample was also reflected by 79.3% of participants indicating a German high school diploma and 16.2% a university degree as their highest educational achievement. Participants were recruited via social networking sites and from courses requiring research participation. Students of allegeable courses were compensated with course credit for their participation.

MEASURES

Remote Associates Test

We presented participants with 20 RAT items. Participants were given 60 seconds to solve each item. Afterwards, participants were either able to check the answer they had entered or—whenever no answer was entered—asked to enter the last word they had considered as the solution. There was no time limit for checking or completing the answer. Because participants were able to review their answer, they were potentially able to change their answer completely (even though we instructed them to only check or complete it). We accounted for this by treating altered responses separately from corrected or completed responses (see below). We selected 10 items from a German CRAT item bank (Landmann et al., 2014). Our scoring should show most promise when most participants fail to answer items correctly and generate a variety of different responses. We speculated that such variance on participants' responses might be somewhat restricted when using only CRAT items. Mednick and Mednick's (1967) instructions for ARAT items merely require the solution to be related to each target word by any kind of association. We assumed that this less restrictive instruction would allow for more variance in the participants' answers. Therefore, we also translated and adapted Mednick and Mednick's (1967) ARAT items into German and again chose ten of those items.

In order to select both CRAT and ARAT items, we pretested 21 ARAT (translations and adaptions of original RAT items from Mednick & Mednick, 1967) and 82 CRAT items (items from the German compound RAT item bank with probabilities of valid solution ranging from 20% to 80%) to five research interns. We applied our proposed scoring to those answers and selected the 10 ARAT and 10 CRAT items with the most variance on our scoring. We only chose items with solution probabilities (as indicated in Landmann et al., 2014) below 60% as we assumed that our scoring should work better for more difficult items.

Standard scoring

As recommended by Mednick and Mednick (1967), all correct answers were scored as 1 and all other answers were scored as 0. All missing answers were treated as incorrect and also scored as 0. For each person, we computed an overall score by summing across all items.

LSA scoring

We used the cosine similarities between each answer and the correct solution in a semantic space as a person's score to the respective RAT item. We used the semantic space *dewak100k_lsa* from the Repository for Semantic Spaces of the Eberhard Karls University Tübingen (http://www.lingexp.uni-tuebingen.de/z2/LSA spaces/). The employed semantic space was created from a term-document matrix with the 100,000 most frequent words from a 800,000 word sdeWaC corpus (http://wacky.sslmit.unibo.it/doku.php?id=corpora). Instead of the common log-entropy weighting scheme, a positive pointwise mutual information weighting scheme was applied to the term-document matrix (Günther, Dudschig & Kaup, 2015). According to Günther et al. (2015), the influence of his choice of weighting scheme on the resulting semantic space is negligible. To reduce the semantic space from 1.5 million to 300 dimensions, singular value decomposition was used. To compute the cosine similarities, we used the R (R Core Team, 2017) package LSAfun (Günther et al., 2015). For each person, we computed an overall score by averaging across the cosine similarities for all items.

Insight problems

We presented participants with three insight problem tasks: the coin problem (Metcalfe, 1986), the egg problem (Sternberg & Davidson, 1982), and the waterlily problem (Schooler, Ohlsson & Brooks, 1993). Each problem was presented on its own and the order in which the problems were presented was randomized. Participants were given three minutes to solve each problem. Solutions were scored with 1 and incorrect (or incomplete) solutions were scored with 0. The total score for each participant was derived by summing across the three problems.

Productive-Reproductive Thinking Inventory

The Productive-Reproductive Thinking Inventory (P-R I; Cunningham & MacGregor, 2016) assesses cognitive styles when solving insight-problems. The questionnaire consists of 21 items. We translated the items into German. A 6-point scale was used (from 1 = *very much atypical for me* to 6 = *very much typical for me*). Separate scores for reproductive, experiential, and normative productive thinking were calculated. In the original paper, insight-problem-solving correlated with reproductive thinking positively (r = .27), with normative productive thinking negatively (r = −.24), and with experiential productive thinking negatively but not significantly so, r = −.14.

Intelligence tasks

Mini-q

We assessed participants' intelligence with the mini-q (Baudson & Preckel, 2015). It allows for an estimate of a person's cognitive abilities in three minutes. Indications for construct validity of the mini-q have been provided: the mini-q correlates among others with the grade point average (r = −.28), with other measures of intelligence (e.g. CFT 20-R: r = .51), and with processing speed (r = .73), but only weakly with task motivation, r = .17.

Spot-the-word test

Participants' verbal intelligence was assessed with a German adaptation of Baddeley, Emslie and Nimmo-Smith's (1993) Spot-the-Word test (SWT). The test consists of 60 word pairs with one word and one pseudo word each. We created 60 German items by selecting fairly to highly uncommon words from the dictionary (http://www.duden.de/definition) and generating corresponding pseudo-words using Wuggy (Keuleers & Brysbaert, 2010). Pseudo-words consisted of the same number of letters as the corresponding item words. All 60 items were presented in randomized order and each item on a separate page. Participants had three-seconds per item. High correlations with the National Adult Reading Test (r = .831–.859) attest to the original test's validity.

Divergent thinking task

Alternate Uses Tasks (AUT; Wallach & Kogan, 1965) were used. The objects in this study were *sphere* and *box* (German: *Kugel* and *Kiste*). Participants were asked to generate alternative uses which were to be as creative as possible in order to assess divergent thinking more validly (Nusbaum, Silvia & Beaty, 2014). Participants' full response sets (Silvia, Martin & Nusbaum, 2009) of ideas to both AUT tasks were rated on a 5-point scale ranging from 1 = *low-quality* to 5 = *high-quality*. Raters were instructed to base their creative quality rating on four dimensions: uncommonness, remoteness, cleverness, and appropriateness. For a description of the first three dimensions, see Silvia et al. (2008). Appropriateness was added in order to be more consistent with common definitions of creativity (e.g., Mednick, 1962). Three raters received rater trainings according to Forthmann et al. (2017). The absolute agreement intra-class correlation for the average quality scores was ICC(2,3) = .90, 95%-CI = [.855, .929]. The number of ideas reflected a person's fluency. Both quality and fluency scores were averaged across both tasks.

Current achievement motivation

We assessed participants' current achievement motivation after receiving the instructions and prior to completing the RAT with the short form of the Questionnaire on Current Motivation (QCM; Freund et al., 2011; Rheinberg, Vollmeyer & Burns, 2001). The short form comprises 12 items along four dimensions (anxiety, challenge, interest, and probability of success). All items were presented in randomized order on one page. Participants indicated their agreement to each item on a 7-point Likert scale ranging from 1 = *not at all* to 7 = *very much*. Evidence for the QCM's validity is provided in Rheinberg et al. (2001).

PROCEDURE

The study was conducted using the online survey platform Unipark (www.unipark.com). Initially, participants were informed about the voluntariness, anonymity, and the approximate duration of the survey. Participants were asked for their permission to use their data. If they did not give their permission, participants were redirected to the debriefing. All others continued with the survey and gave demographic information. Afterwards, the RAT was instructed and two example items were given. Before starting the RAT, the QCM was administered. Participants were then presented with the 20 RAT items in randomized order. After the RAT items, participants completed the other measures in a randomized order. Participants then received an elaborate debriefing explaining in detail the study's intent. Participants who asked for feedback were informed about the total number of correctly solved RAT items.

DATA PREPARATION

All responses to the RAT items were spell-checked and if necessary corrected. All words were adjusted to their singular or infinitive. Responses such as "?", "no idea", or single letters were recoded as missing. Moreover, anything additionally stated within parentheses was deleted. This resulted in four response vectors per item: (a) containing complete answers given within the 60-seconds time limit (in-time responses), (b) incomplete in-time responses that were completed afterwards (completed responses), (c) additionally allowing answers which were altered after the elapse of the 60-seconds time limit (altered responses), and (d) fully untreated responses (raw responses). Compound words which were not contained in the semantic space (e.g., 1.11% of in-time responses) were split up into their components. The random cosine value in the employed semantic space for any pair of words and any single word is higher (M = 0.14) than the random cosine value for any two single words (M = 0.11). To correct for this bias (Forthmann et al., 2018), we subtracted the difference between the two random cosine values from the cosine values of all separated compound words. Any other answers not contained in the employed semantic space were recoded as missing values. All missing values originating from missing answers (e.g., 23.62% of in-time responses) and negative cosine values were recoded as 0 (Günther & Marelli, 2016). As the standard scoring accounts for missing answers by scoring them as 0, we wanted the LSA scoring to do the same. We additionally set negative cosine values to 0 so that no answer was scored less favorably than giving no answer at all. Missing values originating from corpus missings (e.g., 0.95% of in-time responses) remained missing values.

RESULTS

For both scorings, scores for all response types correlated highly with the in-time responses as well as with each other (see Table 1). Thus, we concluded that they offered little different information compared to the in-time responses and therefore we decided to concentrate on the latter for our analyses. Thus, in the following, we are simply going to refer to standard and LSA scores, meaning the standard and LSA scores for the in-time responses.

CLASSIC TEST THEORY ITEM STATISTICS

Item difficulties are displayed separately for CRAT (Table 2) and ARAT items (Table 3). For the CRAT items, the most difficult item in our study was also the most difficult item of the selection according to Landmann et al. (2014). In our study, the mean solution probability across all CRAT items was 0.33 (SD = 0.47). The ARAT items displayed descriptively lower solution probabilities overall, with a mean solution probability across all ARAT items of 0.15 (SD = 0.36). The most difficult ARAT item (Item 19) was in fact solved by none of the participants.

For the LSA scoring, lower mean cosine values indicate higher item difficulty. The rank order of average cosine values for only CRAT items aligns well with the rank order of solution probabilities reported by Landmann et al. (2014) (Kendall's τ = .60). For the ARAT items, rank orders for average cosine values and solution probability were less similar (Kendall's τ = .20). The CRAT items descriptively displayed a higher average cosine (M = 0.46, SD = 0.41) value than the ARAT items (M = 0.33, SD = 0.35), indicating that they were overall slightly easier. However, the difference is less pronounced than the one we found for the solution probabilities. For a graphic display of associations between the average scores, see Figure 1.

TABLE 1. Correlations Between the Different Response Vectors for Both Scorings

	Standard scoring				LSA scoring			
	In-time (1)	Completed (2)	Altered (3)	Raw (4)	In-Time (5)	Completed (6)	Altered (7)	Raw (8)
1	.52							
2	.99***	.55						
3	.98***	1.00***	.57					
4	.97***	.95***	.95***	.50				
5	.88***	.84***	.84***	.85***	.50			
6	.94***	.95***	.95***	.92***	.89***	.56		
7	.94***	.95***	.95***	.92***	.89***	1.00***	.57	
8	.88***	.84***	.84***	.85***	.98***	.88***	.87***	.51

Notes. In-Time = in-time responses. Completed = completed responses. Altered = altered responses. Raw = raw responses (in case of the LSA raw scores, missing values were set to 0). Values across the diagonal represent Cronbach's α for the respective scores. Item 19 had no variance on the standard scoring and was therefore excluded from reliability estimates. ***$p < .001$.

Furthermore, we computed item discriminations for all items as part-whole corrected item-total correlations (see Tables 2 and 3). It is generally desirable for corrected item-total correlations to be greater than .20 (Crocker & Algina, 1986). We also reported 95%-confidence intervals for all part-whole corrected item-total correlations (see Tables 2 and 3) which show that even though the respective correlation in some cases does not surpass .20, the 95%-confidence interval at least includes .20.

For standard scoring part-whole correlations ranged from .01 to .40 with the majority of items below .20. For LSA scoring the range was comparable (from −.05 to .40), although on average the values were slightly lower. For standard scoring of CRAT items only, the values ranged from .10 to .40 (LSA-scoring: from .05 to .38) and for standard scoring of ARAT items values ranged from −.12 to .27 (LSA-scoring: from .01 to .35).

Reliability estimates for the overall RAT standard and LSA score as well as for separate CRAT and ARAT standard and LSA scores are displayed in Table 4. Reliability as indicated by Cronbach's α was modest but comparable for standard and LSA scores. When examining separate CRAT and ARAT scores, reliability estimates were even lower (see Table 4) which can be ascribed to the smaller number of items.

VALIDITY ANALYSES

The correlations for the overall RAT scores and the validity measures are displayed in Table 4. Correlational patterns for separate CRAT and ARAT scores are provided in the Appendix S1. The overall scores correlated highly with each other. The standard scores correlated significantly and positively with the SWT, indicating evidence for convergent validity. Unexpectedly, the cosine-based overall LSA score did not correlate significantly with the SWT. The overall scores positively correlated with insight problems (small effect) and also with the mini-q, however, contrary with our expectations not significantly so. In regard to the PI-R, all overall scores correlated only significantly with the experiential reproductive thinking dimension. Against our expectations, these correlations were positive. Discriminant validity for the overall scores was indicated by the weak and not significant correlations with AUT quality and fluency. Except for the overall score for the standard scoring and the interest dimension, which were significantly and positively associated, none of the overall scores were related to any of the QCM dimensions.

ADDITIONAL ANALYSES: UNSOLVED LSA SCORES

An argument can be made that the LSA scores contain two different types of information: (a) about the ability to correctly solve an item, and (b) the ability to give an answer which is more (or less) semantically similar to the correct solution. The first type of information is shared between LSA and standard scoring, the second is unique to the LSA scoring. We decided post-hoc that when evaluating the LSA scoring, we should also take a look at the information uniquely contained in the LSA scoring. Thus, we excluded all correct answers (i.e., values equal to 1) from the LSA scores and related both an average and item-based unsolved LSA score(s) to the overall standard score as well as the validity measures, finding no significant

TABLE 2. For all CRAT Items, Item Difficulties (as Indicated by Probability of Valid Solution and Mean Cosine Value) and Item Discriminations (as Indicated by Part-Whole Corrected Item-total Correlations; Once for Only the CRAT Items, Once for all Items) with Standard Deviations Indicated in Round Brackets and 95%-Confidence Intervals Indicated in Square Brackets

RAT items	Item content	English translation	Solution probability		M_{LSA} (SD_{LSA})	Standard scoring		LSA scoring (cosines)	
			Present study (SD_{stan})	Landmann et al. (2014)		r_{it} (only CRAT)	r_{it} (all items)	r_{it} (only CRAT)	r_{it} (all items)
RAT 5	Lehrer – Taste – Unterricht (Klavier)	Teacher – key – lesson (piano)	0.13 (0.33)	0.23	0.23 (0.33)	.25 [.06, .41]	.19 [.00, .36]	.15 [−.04, .33]	.06 [−.13, .24]
RAT 2	Deutung – Note – Tänzer (Traum)	Interpretation – key – dancer; (dream)	0.16 (0.37)	0.58	0.33 (0.32)	.18 [−.01, .35]	.21 [.02, .38]	.17 [–02, .35]	.19 [0, .36]
RAT 1	Wurm – Haar – Scheibe (Band)	Worm – hair – disc (tape)	0.22 (0.41)	0.40	0.29 (0.38)	.11 [−.08, .29]	.11 [−.08, .29]	.13 [−.05, .31]	.13 [−.06, .31]
RAT 10	Natur – Tüte – Welt (Wunder)	Nature – bag – world (wonder)	0.25 (0.43)	0.45	0.41 (0.36)	.17 [−.02, .35]	.08 [−.10, .27]	.17 [−.01, .35]	.12 [−.07, .30]
RAT 9	Wirtschaft – Tür – Eingang (Haus)	Economy – door – entrance (house)	0.26 (0.44)	0.38	0.38 (0.39)	.21 [.02, .38]	.20 [.01, .37]	.17 [−.01, .35]	.12 [−.07, .30]
RAT 6	Spiegel – Qualität – Röhre (Bild)	Mirror – quality – tube (image)	0.39 (0.49)	0.55	0.51 (0.41)	.40 [.23, .54]	.40 [.23, .55]	.38 [.20, .52]	.40 [.23, .54]
RAT 4	Streit – Bruch – Versprechen (Ehe)	Fight – break – promise (marriage)	0.41 (0.49)	0.58	0.58 (0.38)	.23 [.04, .40]	.24 [.06, .41]	.21 [.03, .39]	.16 [−.03, .34]
RAT 7	Angel – Motor – Profi (Sport)	Fishing rod – motor – pro (sports)	0.43 (0.50)	0.35	0.51 (0.44)	.23 [.04, .40]	.18 [.00, .36]	.19 [0, .36]	.16 [−.03, .34]
RAT 3	Werk – Alarm – Leiter (Feuer)	Factory – alarm – ladder (fire)	0.54 (0.50)	0.38	0.66 (0.39)	.13 [−.05, .31]	.14 [−.05, .32]	.05 [−.14, .24]	.10 [−.09, .28]
RAT 8	Hütte – Steuer – Futter (Hund)	Hut – tax – fodder (dog)	0.56 (0.50)	0.48	0.65 (0.43)	.10 [−.09, .28]	.20 [.02, .37]	.05 [−.14, .24]	.14 [−.05, .32]

Note. SD_{stan} = standard deviation of standard scores. M_{LSA} = average cosine values (LSA scoring). SD_{LSA} = standard deviation of LSA scores. r_{it} (only CRAT) = part-whole corrected item-total correlations, calculated only across CRAT items. r_{it} (all items) = part-whole corrected item-total correlations, calculated across all items. Solutions are in parentheses. English translations are displayed for information purposes only and do not necessarily represent functioning items in English.

TABLE 3. For All ARAT Items, Item Difficulties (as Indicated by Probability of Valid Solution and Mean Cosine Value) and Item Discriminations (as Indicated by Part-whole Corrected Item-total Correlations; Once for Only the ARAT Items, Once for All Items) with Standard Deviations Indicated in Round Brackets and 95%-confidence Intervals Indicated in Square Brackets

RAT items	Item content	English translation	Solution probability (SD_{stan})	M_{LSA} (SD_{LSA})	Standard scoring		LSA scoring (cosines)	
					r_{it} (only ARAT)	r_{it} (all items)	r_{it} (only ARAT)	r_{it} (all items)
RAT 19	Kirsche – Zeit – Duft (blühen)	Cherry – time – scent (flourish)	0.00 (0.00)	0.46 (0.32)	–	–	.19 [0, .36]	.28 [.10, .44]
RAT 20	Beilage – Niederlage – Taschentuch (Packung)	Garnish – defeat – tissue (box)	0.01 (0.09)	0.15 (0.17)	−.12 [−.30, .07]	.01 [−.18, .19]	.03 [−.16, .21]	−.04 [−.23, .14]
RAT 13	Elefant – Lücke – lebhaft (Gedächtnis)	Elephant – gap – lively (memory)	0.03 (0.16)	0.18 (0.22)	.16 [−.03, .34]	.08 [−.11, .26]	.18 [0, .36]	.03 [−.16, .22]
RAT 17	Maus – streng – blau (Käse)	Mouse – strong – blue (cheese)	0.04 (0.19)	0.11 (0.20)	−.01 [−.19, .19]	.01 [−.17, .20]	.01 [−.18, .20]	−.05 [−.24, .14]
RAT 16	Herrscher – kreischen – Wappen (Adler)	Sovereign – screech – coat of arms (eagle)	0.09 (0.29)	0.36 (0.27)	.01 [−.17, .20]	.01 [−.19, .19]	.15 [−.04, .33]	.03 [−.15, .22]
RAT 18	Bier – Schmetterling – Entscheidung (Bauch)	Beer – butterfly – decision (stomach)	0.14 (0.34)	0.28 (0.32)	.16 [−.03, .34]	.22 [.03, .39]	.15 [−.04, .32]	.17 [−.02, .35]
RAT 12	Bummeln – putzen – Aussicht (Fenster)	Stroll – clean – view (window)	0.20 (0.40)	0.36 (0.36)	.26 [.08, .43]	.21 [.02, .38]	.23 [.04, .40]	.19 [0, .37]
RAT 11	Bass – komplex – schlafen (tief)	Bass – complex – sleep (deep)	0.23 (0.42)	0.41 (0.37)	.18 [−.01, .35]	.08 [−.11, .26]	.35 [.17, .50]	.18 [0, .36]
RAT 15	Eifersucht – Golf – Bohnen (grün)	Jealousy – golf – beans (green)	0.36 (0.48)	0.44 (0.44)	.10 [−.09, .28]	.06 [−.12, .25]	.15 [−.04, .33]	.13 [−.06, .31]
RAT 14	Humor – Magie – Tod (schwarz)	Humour – magic – death (black)	0.46 (0.50)	0.59 (0.41)	.27 [.08, .43]	.38 [.21, .53]	.19 [0, .37]	.33 [.15, .49]

Note. SD_{stan} = standard deviation of standard scores. M_{LSA} = average cosine values (LSA scoring). SD_{LSA} = standard deviation of LSA scores. r_{it} (only ARAT) = part-whole corrected item-total correlations, calculated only across ARAT items. r_{it} (all items) = part-whole corrected item-total correlations, calculated across all items. Item content = respective RAT item with the correct solution and the English translation each in parentheses. Items were only presented in German. English translations are displayed for information purposes only and do not necessarily represent functioning items in English. No part-whole corrected item-total correlations were computed for the standard scoring for RAT item19 as there was no variance on this item when scored with the standard scoring.

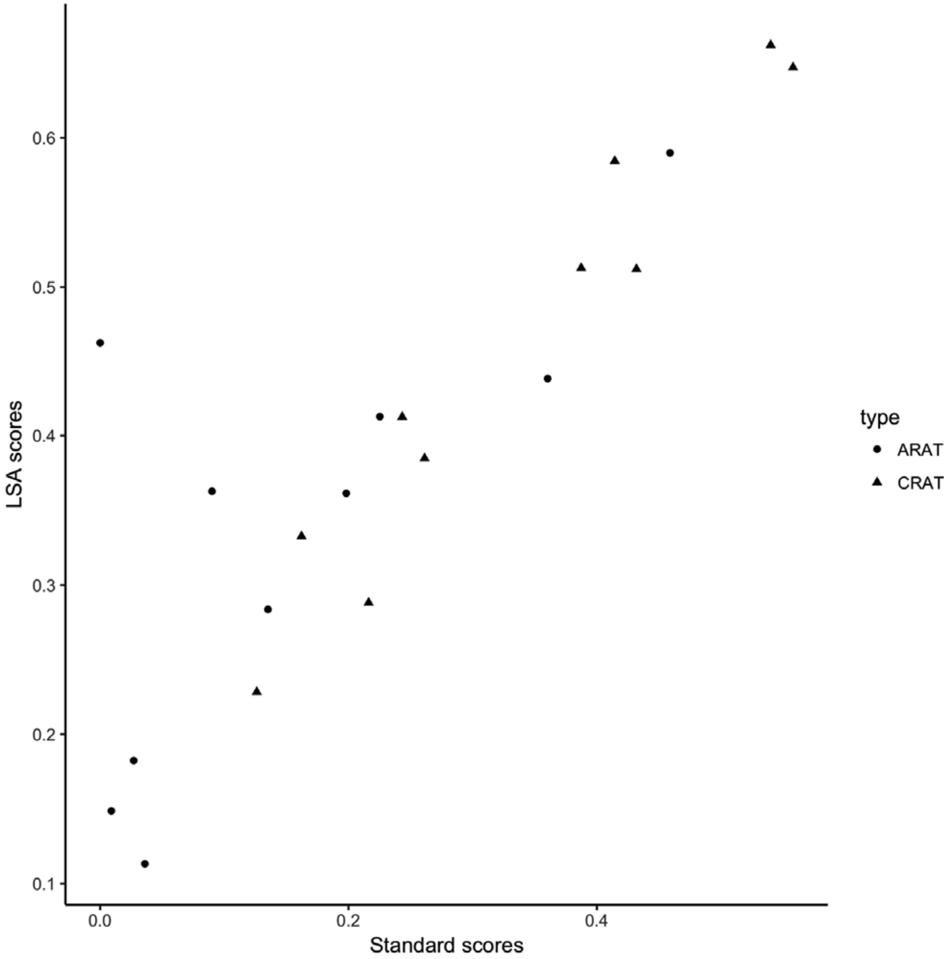

FIGURE 1. Relation between average standard (*x*-axis) and LSA (*y*-axis) scores. Dots represent associational Remote Associates Test (ARAT) items, triangles represent compound Remote Associates Test (CRAT) items.

correlations except for with three out of the four motivation dimensions (see the Appendix S1 for the correlations as well as a more detailed description of our analyses). For the three motivation dimensions, the correlations were negative, indicating that the more motivated participants were, the lower their unsolved LSA scores were.

DISCUSSION

The aim of this study was to evaluate an alternate scoring based on LSA for the Remote Associates Test (RAT; Mednick, 1962; Mednick & Mednick, 1967). In addition, we investigated and compared the psychometric properties of the LSA scoring to those of the standard scoring. Validity for both scorings was evaluated by relating them to measures of convergent and divergent thinking.

PSYCHOMETRIC PROPERTIES OF THE STANDARD AND LSA SCORING

As intended, the items were rather difficult with no solution probability exceeding 60%. For CRAT items, the order of difficulties for the LSA scoring was comparable to standard scoring. However, neither

TABLE 4. Correlations of RAT Scores With Validity (Convergent and Discriminate) Criterion Measures

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Standard (overall)	**.52**													
2. LSA cosine (overall)	.88***	**.50**												
3. Mini-q	.18	.17	**.91**											
4. SWT	.24*	.17	.47***	**.73**										
5. AUT (Q)	.11	.10	.13	.19	**.44**									
6. AUT (F)	.03	.09	.09	−.02	.42***	**.77**								
7. Insight problem	.17	.13	.23*	.12	.16	.21*	**.48**							
8. P-R I (Pr)	.09	.03	.09	.04	.07	.11	.25**	**.92**						
9. P-R I (Ex)	.21*	.19*	−.11	.13	.01	−.13	−.02	−.38***	**.89**					
10. P-R I (No)	.18	.17	−.12	.01	−.12	−.20*	−.08	−.45***	.78***	**.82**				
11. QCM (PoS)	−.13	−.03	.11	.04	.01	−.01	.20*	.22*	.00	−.06	**.71**			
12. QCM (A)	.11	−.04	−.02	.13	.01	−.03	.16	.01	.08	.13	−.49***	**.87**		
13. QCM (I)	.21*	.05	.10	.03	.17	.02	.22*	.32***	.02	−.09	.12	.11	**.77**	
14. QCM (C)	.15	.01	.13	.24*	.11	−.10	.25**	.27**	.10	−.02	−.04	.27**	.43***	**.70**

Notes. Standard (overall) = overall RAT scores based on the standard scoring. LSA cosine (overall) = overall RAT scores based on the LSA cosine values. LSA cosine (CRAT) = overall RAT scores for CRAT items only based on the LSA cosine values. Mini-q = Mini-q test scores. SWT = Spot-the-Word test scores. AUT (Q) = quality score from the Alternate Uses Task. AUT (F) = fluency score from the Alternate Uses Task. Insight problems = sum scores from the three insight problems. P-R I (Pr) = Productive-Reproductive Inventory, productive thinking dimension. P-R I (Ex) = Productive-Reproductive Inventory, experiential reproductive thinking dimension. P-R I (No) = Productive-Reproductive Inventory, normative reproductive thinking dimension. QCM (PoS) = Questionnaire on Current Motivation, probability of success dimension. QCM (A) = Questionnaire on Current Motivation, anxiety dimension. QCM (I) = Questionnaire on Current Motivation, interest dimension. QCM (C) = Questionnaire on Current Motivation, challenge dimension. Cronbach's α for all measures is displayed across the diagonal in boldface. *$p < .05$, **$p < .01$, ***$p < .001$.

were well aligned with the order of difficulties found by Landmann et al. (2014). In line with Hung, Huang and Chen (2016), but in contrast to Olzmann (2012), the ARAT items tended to be more difficult than the CRAT items. Moreover, for the ARAT items, the orders of item difficulties for the standard and the LSA scoring were not as comparable as for the CRAT items. The discrepancies foremost occurred for high difficulty ARAT items, suggesting that for those items, the LSA scoring offered additional information to standard scoring (most obvious for item 19).

We found rather small item discriminations in general—more so for ARAT than for CRAT items and for LSA than for standard scoring. Some items within each group correlated higher with the overall score than with the scores of their own type, indicating that in each case, the item tended to be more representative of the respective other item type (for example, item 18). Another interesting observation was that item 19, which had to be excluded from standard scoring analyses due to zero variance, was one of the few ARAT items which—when scored with the LSA scoring—displayed satisfactory discrimination. In future studies, a larger item pool should be used, to identify a larger number of items of overall good psychometric properties. As we started off with only 10 items per item type, this was not a viable option for this study.

VALIDITY EVIDENCE FOR THE STANDARD AND LSA SCORING

The correlational patterns we found indicated that LSA scoring did in fact demonstrate comparable validity to standard scoring. In accordance with previous research (*r*'s ranging from .13 to .07 in Akbari Chermahini et al., 2012; *r*'s from −.05 to .13 in Lee et al., 2014), overall RAT scores in this study correlated weakly with measures of divergent thinking (*r*'s from −.01 to .13) which is usually interpreted as evidence for discriminant validity for the RAT. Results regarding convergent validity were less unambiguous and tended to be less strong than patterns found previously. For instance, except for the overall standard RAT scores and the vocabulary task (*r* = .24), we found none of the expected associations between overall RAT scores and convergent thinking. Even though standard RAT scores have previously been found to correlate moderately with fluid intelligence (*r* = .47 in Akbari Chermahini et al., 2012; *r*'s from .33 to .34 in Lee et al., 2014) and with vocabulary tasks (*r*'s from .41 to .42 in Lee et al., 2014), we only found weak and non-significant correlations. However, validity evidence was consistent across standard and LSA scoring and thus does not indicate specifically poor convergent validity for the LSA scoring. Moreover, correlations of overall RAT scores and performance on insight problems (*r*'s from .13 to .17) were of comparable size to the ones found in Lee et al. (2014; *r*'s from .14 to .18) who used a bigger sample, indicating that the effects found in our study might have been significant in a larger sample as well. The weak to lacking relationships of both scorings to the constructs included in our validity analyses should always be judged in the light of what degree of association we should expect for these constructs. Even in previous studies examining the validity of the RAT (Akbari Chermahini et al., 2012; Lee et al., 2014), reported associations were moderate at best. The constructs themselves may arguably be only distally related to RAT performance. The unexpected lack of differences in validity for both scorings may indicate that the standard scoring already contains most of the information that there is in performance differences. Thus, the most important advantage of the LSA scoring is its ability to give information about performance differences about very difficult RAT items which are unsolvable in most samples. We elaborate on this idea below.

In line with our expectations, the overall standard RAT scores were positively related to the interest facet of current motivation, however, the overall RAT scores based on the LSA scoring were not. Contrary with expectations based on previous research (Cunningham & MacGregor, 2016), we did not find a similar correlational pattern for the overall RAT scores than for insight problems when relating both to measures of cognitive styles employed when solving problems. In fact, the only cognitive style we found to be (positively) related to the overall RAT scores was experiential reproductive thinking which is a cognitive style that previous research has found to be—at least as a trend—negatively related to performance on insight problems (Cunningham & MacGregor, 2016).

Our post-hoc analyses using unsolved LSA scores (i.e., LSA scores from which all correct answers were excluded) which were meant to separate information shared by both scorings (i.e., the ability to answer an item correctly) from information unique to the LSA scoring (i.e., the ability to give an answer which is more or less semantically similar to the correct solution) revealed no relevant associations to any of the constructs selected to evaluate validity, except for negative relations with motivation. This finding might be explained by the problems associated with the approach of excluding solved trials for the LSA-scoring which we discuss in more detail in the Appendix S1. In short, we confound sampling error of item scores with item difficulty and reliability of average LSA-unsolved scores with ability. In addition, excluding information about participants' ability which should be reflected in our scoring does not seem reasonable. A more promising approach to investigate the same question may be to use extremely difficult RAT items in samples where they are unsolvable. Thus, we would not have to eliminate relevant information from the scores but could still evaluate the information unique to the LSA scoring. In fact, under such conditions, the standard scoring would not offer any information at all, resulting in zero variance and making any psychometric analyses impossible. As our LSA-scored items differed greatly in range (without 1), we advise researchers to investigate this question to use RAT items with large cosine ranges.

An interesting exploratory finding of this study is that—as reported in the Appendix S1—correlational patterns for separate CRAT and ARAT scores diverged from each other. This finding deserves further attention and an attempt at replication in future research as the diverging correlational pattern as well as low inter-correlations between CRAT and ARAT scores point to different underlying abilities (see also Worthen & Clark, 1971).

IMPLICATIONS FOR THE LSA SCORING

Judging from the results of our validity analyses, the LSA seems to assess the same processes as the standard scoring which is very promising. Moreover, an apparent advantage of the LSA over the standard scoring was demonstrated in our study by item 19: Any items for which generating the correct answer is very difficult run the danger of showing almost no or zero-variance on the standard scoring. Such items have to be excluded from any psychometric analyses when using the standard scoring. LSA scoring, however, assigns different values for different responses and items which no participant solved correctly pose no such problem. Generally, LSA scoring seems more promising when more difficult RAT items are used. Future research could use the LSA scoring to examine the psychometric properties of RAT items which are unsolvable in most samples and thus could not be examined using the standard scoring.

In relation to this, the potential loss of information for the standard scoring appeared to be less dramatic than expected. For example, higher standard deviations for LSA scores were only observed for items with solving probability below 10%. For easier items it was even the case that more variation for the standard scoring was observed. Thus, LSA scoring seems to be most promising when large item pools are needed and a choice of only moderately hard items is not possible. Olteţeanu, Schultheis and Dyer (2018), for example, presented an automatic RAT item-generator and some of their items had very low solving probabilities. If such item-generators are used to generate items on the fly in online or computerized testing, LSA scores are likely to be informative beyond standard scoring.

FURTHER IMPLICATIONS

Our findings are also relevant to the P-R I (Cunningham & MacGregor, 2016). So far, the questionnaire has only been validated by relating the two dimensions (i.e., productive and reproductive thinking) to insight problem performance, finding the reproductive dimension to be positively, and the productive dimensions to be negatively related to insight-problem-solving. While those patterns were mostly replicated in our sample, we found very different patterns for the RAT scores as we would have expected based on the assumption that insight problems and RAT items assess similar cognitive processes (Bowden & Jung-Beeman, 2003). The only dimension that we found to be related to RAT performance was experiential reproductive thinking which has been inversely linked with performance on insight problems (Cunningham & MacGregor, 2016). However, in our study, we found experiential reproductive thinking to be positively related to both RAT scores. The latter association in particular raises doubts as to how comparable insight problems and RAT items are.

LIMITATIONS

Any implications derived from our findings should always be considered carefully in light of the limitations of this study. For example, a larger and more diverse sample would allow for better generalizability of the results. Any problems common to LSA represent limitations to this study. For example, not all answers given by participants are contained in the semantic space; such corpus missing values are inevitable. Depending on the LSA-provided semantic space from which measures of semantic associations (i.e., cosine values) are calculated, such measures may take on negative values which are hard to interpret and therefore often set to 0 (Günther & Marelli, 2016). This may not be an ideal solution. A more practical limitation of LSA scoring is the extensive data preparation our data underwent in order to obtain the LSA scores. However, the high correlations between the raw LSA scores and LSA scores based on the prepared data indicated that such extensive data preparation might not be necessary.

CONCLUSION

Apart from some discrepancies, we found similar correlational patterns for the alternate LSA scoring and the standard RAT scoring when evaluating the validity of both scorings. Thus, we concluded that the alternate LSA scoring showed comparable validity in relation to the standard RAT scoring. This suggests that the standard dichotomous RAT scoring already contains a great amount of information about individual differences in performance. Finally, several methodological issues were identified to further explore LSA-based scoring of RAT items.

REFERENCES

Akbari Chermahini, S., Hickendorff, M., & Hommel, B. (2012). Development and validity of a Dutch version of the Remote Associates Task: An item-response theory approach. *Thinking Skills and Creativity*, *7*, 177–186. https://doi.org/10.1016/j.tsc.2012.02.003.

Ansburg, P.I. (2000). Individual differences in problem solving via insight. *Current Psychology*, *19*, 143–146. https://doi.org/10.1007/s12144-000-1011-y.

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The Spot-the-Word test: A robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology*, *32*, 55–65. https://doi.org/10.1111/j.2044-8260.1993.tb01027.x.

Balas, R., Sweklej, J., Pochwatko, G., & Godlewska, M. (2012). On the influence of affective states on intuitive coherence judgements. *Cognition and Emotion*, *26*, 312–320. https://doi.org/10.1080/02699931.2011.568050.

Baudson, T.G., & Preckel, F. (2015). mini-q: Intelligenzscreening in drei Minuten. *Diagnostica*, *62*, 182–197. https://doi.org/10.1026/0012-1924/a000150.

Bolte, A., & Goschke, T. (2005). On the speed of intuition: Intuitive judgments of semantic coherence under different response deadlines. *Memory and Cognition*, *33*, 1248–1255. https://doi.org/10.3758/bf03193226.

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

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

Chun, C.A., & Hupé, J.M. (2016). Are synesthetes exceptional beyond their synesthetic associations? A systematic comparison of creativity, personality, cognition, and mental imagery in synesthetes and controls. *British Journal of Psychology*, *107*, 397–418. https://doi.org/10.1111/bjop.12146.

Crocker, L. S., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.

Cunningham, J. B., & MacGregor, J. N. (2016). A self-report measure of productive thinking in solving insight problems. *The Journal of Creative Behavior*; Advance online publication. https://doi.org/10.1002/jocb.169.

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

Forthmann, B., Holling, H., Zandi, N., Gerwig, A., Çelik, P., Storme, M., & Lubart, T. (2017). Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores. *Thinking Skills and Creativity*, *23*, 129–139. https://doi.org/10.1016/j.tsc.2016.12.005.

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2018). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior*; Advance online publication. https://doi.org/10.1002/jocb.240.

Freund, P.A., Kuhn, J.T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*, 629–634. https://doi.org/10.1016/j.paid.2011.05.033.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun – An R package for computations based on latent semantic analysis. *Behavior Research Methods*, *47*, 930–944. https://doi.org/10.3758/s13428-014-0529-0.

Günther, F., & Marelli, M. (2016). Understanding karma police: The perceived plausibility of noun compounds as predicted by distributional models of semantic representation. *PLoS ONE*, *11*, e0163200. https://doi.org/10.1371/journal.pone.0163200.

Gupta, N., Jang, Y., Mednick, S.C., & Huber, D.E. (2012). The road not taken: Creative solutions require avoidance of high-frequency responses. *Psychological Science*, *23*, 288–294. https://doi.org/10.1177/0956797611429710.

Hung, S.P., Huang, P.S., & Chen, H.C. (2016). Cognitive complexity in the Remote Association Test-Chinese version. *Creativity Research Journal*, *28*, 442–449. https://doi.org/10.1080/10400419.2016.1229988.

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. https://doi.org/10.3389/fpsyg.2017.00099.

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

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*, 627–633. https://doi.org/10.3758/BRM.42.3.627.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240. https://doi.org/10.1037/0033-295X.104.2.211.

Landmann, N., Kuhn, M., Piosczyk, H., Feige, B., Riemann, D., & Nissen, C. (2014). Entwicklung von 130 deutschsprachigen Compound Remote Associate (CRA)-Worträtseln zur Untersuchung kreativer Prozesse im deutschen Sprachraum. *Psychologische Rundschau*, *65*, 200–211. https://doi.org/10.1026/0033-3042/a000223.

Laughlin, P.R., Kerr, N.L., Munch, M.M., & Haggarty, C.A. (1976). Social decision schemes of the same four-person groups on two different intellective tasks. *Journal of Personality and Social Psychology*, *33*, 80–88. https://doi.org/10.1037/0022-3514.33.1.80.

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*, 446–460. https://doi.org/10.1037/a0036773.

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

Mednick, S. A., & Mednick, M. T. (1967). *Examiner's manual, Remote Associates Test: College and adult forms 1 and 2*. Boston: Houghton Mifflin.

Mednick, M.T., Mednick, S.A., & Mednick, E.V. (1964). Incubation of creative performance and specific associative priming. *The Journal of Abnormal and Social Psychology*, *69*, 84–88. https://doi.org/10.1037/h0045994.

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

Nusbaum, E.C., Silvia, P.J., & Beaty, R.E. (2014). Ready, set, create: What instructing people to "Be Creative" reveals about the meaning and mechanisms of divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *8*, 423–432. https://doi.org/10.1037/a0036549.

Olteţeanu, A.-M., & Falomir, Z. (2015). comRAT-C: A computational compound Remote Associates Test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, *67*, 81–90. https://doi.org/10.1016/j.patrec.2015.05.015.

Olteţeanu, A. M., & Schultheis, H. (2017). What determines creative association? Revealing two factors which separately influence the creative process when solving the Remote Associates Test. *The Journal of Creative Behavior*; Advance online publication. https://doi.org/10.1002/jocb.177.

Olteeanu, A.-M., Schultheis, H., & Dyer, J. (2018). Computationally constructing a repository of compound remote associates test items in American English with comRAT-G. *Behavior Research Methods*, *50*, 1971–1980. https://doi.org/10.3758/s13428-017-0965-8.

Olzmann, A. E. (2012). Problem solving and memory: Investigating the solvability and memorability of remote associates problems (Bachelor thesis, University of Michigan). Retrieved from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/91798/aolzmann.pdf?sequence=1 [last accessed 20.12.2018].

Özen, G., Doğan, A., & Cinan, S. (2015). Uzak Bağlantılar Testi: Norm ve Güvenirlik Çalışması. *Psikoloji Çalışmaları Dergisi*, *35*, 25–46.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rheinberg, F., Vollmeyer, R., & Burns, B.D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen (Langversion, 2001). *Diagnostica*, *2*, 57–66. https://doi.org/10.1026//0012-1924.47.2.57.

Salvi, C., Costantini, G., Bricolo, E., Perugini, M., & Beeman, M. (2016). Validation of Italian rebus puzzles and compound remote associate problems. *Behavior Research Methods*, *48*, 664–685. https://doi.org/10.3758/s13428-015-0597-9.

Schooler, J.W., & Melcher, J. (1995). The ineffability of insight. In S.M. Smith, T.B. Ward & R.A. Finke (Eds.), *The creative cognition approach* (pp. 249–268). Cambridge, MA: MIT Press.

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

Silvia, P.J., Martin, C., & Nusbaum, E.C. (2009). A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. *Thinking Skills and Creativity*, *4*, 79–85. https://doi.org/10.1016/j.tsc.2009.06.005.

Silvia, P.J., Winterstein, B.P., Willse, J.T., Barona, C.M., Cram, J.T., Hess, K.I., . . . & Richard, C.A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 68–85. https://doi.org/10.1037/1931-3896.2.2.68.

Smith, K.A., Huber, D.E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*, 64–75. https://doi.org/10.1016/j.cognition.2013.03.001.

Smith, K.A., & Vul, E. (2015). The role of sequential dependence in creative semantic search. *Topics in Cognitive Science*, *7*, 543–546. https://doi.org/10.1111/tops.12152.

Sternberg, R.J., & Davidson, J.E. (1982). The mind of the puzzler. *Psychology Today*, *16*(6), 37–44.

Wallach, M.A., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity-intelligence distinction*. New York: Holt, Rinehart, & Winston.

Worthen, B.R., & Clark, P.M. (1971). Toward an improved measure of remote associational ability. *Journal of Educational Measurement*, *8*, 113–123. https://doi.org/10.1111/j.1745-3984.1971.tb00914.x.

Marie Beisemann, Boris Forthmann, Paul-Christian Bürkner and Heinz Holling, Westfälische Wilhelms-Universität Münster

Correspondence concerning this article should be addressed to Boris Forthmann, Institut für Psychologie, Westfälische Wilhelms-Universität, Fliednerstrasse 21, 48149, Münster, Germany. E-mail: boris.forthmann@wwu.de

AUTHOR NOTE

Marie Beisemann, Boris Forthmann, Paul-Christian Bürkner, and Heinz Holling, Institut für Psychologie, Westfläische Wilhelms-Universität, Münster, Germany.

This research was supported by grant HO 1286/11-1 of the German Research Foundation (DFG) to Heinz Holling.

We thank Franca Kruppa and Thi Quynh-Nhu Tran for the coding of the divergent-thinking idea sets. In addition, we thank Kristina van Vorst, Annika Weiermann, Julia Scheuss, Franca Kruppa, and Paula Blumenschein for providing RAT responses in a little pilot study. We further thank two anonymous reviewers for their valuable comments on our initially submitted manuscript.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. The appendix includes a subsection on an analysis of unsolved RAT-scores, and a subsection on an exploratory analysis that compares ARAT and CRAT items.