Behav Res (2018) 50:1971–1980
https://doi.org/10.3758/s13428-017-0965-8

Computationally constructing a repository of compound remote associates test items in American English with comRAT-G

Ana-Maria Olteţeanu[1] · Holger Schultheis[1] · Jonathan B. Dyer[2]

Published online: 12 December 2017

Abstract The Remote Associates Test (RAT) has been used to measure creativity, however few repositories or standardizations of test items exist, like the normative data on 144 items provided by Bowden and Jung-Beeman. comRAT is a computational solver which has been used to solve the compound RAT in linguistic and visual forms, showing correlation to human performance over the normative data provided by Bowden and Jung-Beeman. This paper describes using a variant of comRAT, comRAT-G, to generate and construct a repository of compound RAT items for use in the cognitive psychology and cognitive modeling community. Around 17 million compound Remote Associates Test items are created from nouns alone, aiming to provide control over (i) frequency of occurrence of query items, (ii) answer items, (iii) the probability of coming up with an answer, (iv) keeping one or more query items constant and (v) keeping the answer constant. Queries produced by comRAT-G are evaluated in a study in comparison with queries from the normative dataset of Bowden and Jung-Beeman, showing that comRAT-G queries are similar to the established query set.

Keywords Remote associates test · Creative cognition · Computational creativity · Cognitive modeling

✉ Ana-Maria Olteţeanu
amoodu@informatik.uni-bremen.de

1 Cognitive Systems, Bremen Spatial Cognition Center, University of Bremen, Enrique-Schmidt-Straße 5, 28359 Bremen, Germany

2 Department of Computer Science, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

Introduction

Imagine you set out to test or manipulate the creative ability of some of your participants using the Remote Associates Test (Mednick & Mednick, 1971). This test consists of giving participants three words—like WATER, MINE, and SHAKER—and asking them what word might be related to all three such words.[1] You would have to first find and select a set of RAT queries, and then control as many variables as you can about them.

Various creativity tests are available: the Alternative Uses Test (Guilford, 1967), the Remote Associates Test (Mednick & Mednick, 1971), the Torrance Tests of Creative Thinking (Kim, 2006), the Wallach–Kogan tests (Wallach & Kogan, 1965), riddles (used by Whitt & Prentice, 1977; Qiu et al., 2008), empirical insight tests from Duncker (1945), Maier (1931), and Saugstad and Raaheim (1957) and others. However, some of these tests are not easily available (TTCT), some are non-standardized or do not provide any norms, and others provide small sets of stimuli. Thus, various creativity tests could highly benefit from modernization by being normed, having more factors controlled for and by developing ampler sets of stimuli, in order to provide more varied testing conditions.

The Remote Associates Test (Mednick & Mednick, 1971), has been used to measure creativity and adapted to various other languages—for example in Japanese (Baba, 1982) and Dutch (Chermahini, Hickendorff, & Hommel, 2012). The test has been rated as the second most used creativity test in a meta-analysis surveying 45 neuroimaging studies (Arden, Chavez, Grazioplene, & Jung, 2010). The RAT is assumed to measure creative convergent thinking,

[1] The answer to this RAT query is SALT.

unlike other creativity tests, which are better suited to measure divergent thinking—e.g., the Alternative Uses Test. The RAT is very useful in measuring insight theory effects, as performance in the RAT has been shown to correlate with performance in insight problems (Schooler & Melcher, 1995).

Various types of investigations in and beyond creative cognition use the Remote Associates Test. Amongst others, these include the study of the effects of incubation with GO players (Sio & Rudowicz, 2007) and baseball players (Wiley, 1998), the relation between REM sleep and creativity (Cai, Mednick, Harrison, Kanady, & Mednick, 2009), synesthesia and creativity (Sitton & Pierce, 2004; Ward, Thompson-Lake, Ely, & Kaminski, 2008), the role of affect in problem-solving (Fodor, 1999), memory (Storm, Angello, & Bjork, 2011), peripheral attention (Ansburg & Hill, 2003) etc. Because of its wide use, the scientific community would thus benefit from an ampler set of standardized stimuli for the RAT.

Worthen and Clark (1971) made the case that different stimuli categories can be distinguished within the original stimuli of Mednick and Mednick (1971)—specifically they differentiated between functional items and structural items (sometimes also appearing in the literature under the name of *compound* items).

The CreaCogs framework (Olteţeanu, 2014, 2016) for creative problem-solving uses knowledge organization to support a unified set of core creative problem solving processes, like association, associative convergence, re-representation, restructuring, search, and substitution. The validation of the framework and processes is done in a comparative manner to human performance (Olteţeanu, Falomir, & Freksa, in press) through implementing systems which can show creative problem-solving abilities and solve creativity tests for humans. Among such systems, the comRAT-C system (Olteţeanu & Falomir, 2015) has explicitly addressed the computational solving of compound RAT queries. When solving the RAT, comRAT-C calculated the probability of finding a solution based on the frequency of query and answer words, as will be shown in "Generating new remote associates test items with comRAT-G". A highly significant correlation has been observed between the results of comRAT-C and the difficulty of RAT queries for humans, as expressed in percentage of solvers and response times in the human normative data (Bowden & Jung-Beeman, 2003). This correlation ranged between 0.3 and 0.52 for different solving times. This correlation showed that the frequency of query items plays an important role into how difficult a query is for human participants, and a standardized set of RAT stimuli should take this into account. Furthermore, (Olteţeanu & Schultheis, in press) modified frequency and probability factors independently, keeping them at low and high levels. This showed that frequency and probability both are factors which influence accuracy and response times when solving the RAT.

For the Remote Associates Test, normative data from human participants does exist from Bowden and Jung-Beeman (2003), which provides data on mean time-to-solution and percentage of participants solving for 144 compound RAT problems, with four different time limits. Though very useful, this work does not provide standardized queries based on frequency of occurrence of word or answer stimuli, the importance of which has been shown by Olteţeanu and Falomir (2015). This paper aims to enrich the existing pool of compound Remote Associates Test items and provide a standardized treatment that allows control over the frequency of occurrence and probability of finding an answer. Seventeen million new compound RAT items are constructed, using the entire space of frequent noun expressions in American English - thus providing the largest standardized treatment of the compound RAT test to date. These items are computationally generated by adapting the previously implemented computational solver of the Remote Associates Test, comRAT-C (Olteţeanu & Falomir, 2015), to a generative variant - comRAT-G. The frequency of items from the COCA corpus[2] and comRAT-C's probability of finding an answer are indexed in the provided repository, and can be used to generate a set of controlled queries. Subsets of queries in which one word or the answer are kept constant can also be extracted.

The rest of this paper is structured as follows. A brief overview of the comRAT solver and its transformation to comRAT-G is provided in "From comRAT to comRAT-G". The methodology of generating new Remote Associates Test items with comRAT-G is explained, together with examples, in "Generating new remote associates test items with comRAT-G". The evaluation of 100 queries with human participants is shown in "Evaluation". The type of data generated in the repository is described in "The Repository", and various possible uses are showcased. Finally, future work is discussed. The 100 items used for evaluation are presented in the Annex.

From comRAT to comRAT-G

The comRAT-C (Olteţeanu & Falomir, 2015) system has been built to solve the compound variant of the

[2]Corpus of Contemporary American English (COCA): http://corpus.byu.edu/coca/

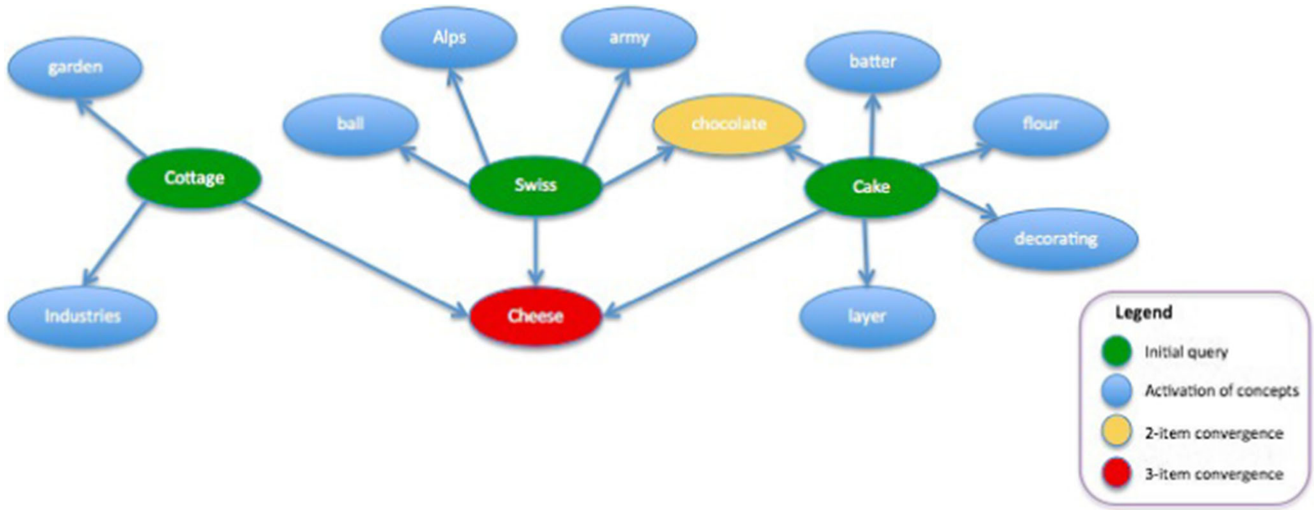

Fig. 1 A graphical depiction of the link structure obtained in comRAT-C. Only a few links are shown for visual readability

Remote Associates Test, in the tradition of Psychometric AI (Bringsjord, 2011) and as an exploration of the processes of the CreaCogs creative problem-solving framework (Olteţeanu, 2014; Olteţeanu & Falomir, 2016).

As data, comRAT-C takes the most frequent 2-grams from the COCA corpus. Knowledge organization plays an important role in CreaCogs and in comRAT's knowledge base in the following way: whenever a 2-gram is given to comRAT-C, it is stored as an object of the Expression class, which is constructed from two objects of the Concept class and a Link between them. If one of the Concepts is already present, the other Concept and the new Link are added. If both Concepts are present, only the Link is added. For example, if the 2-gram *"cake flour"* is read, comRAT-C will check to see if it knows the Concepts *"cake"* and *"flour"*. If it doesn't know one of them, the item will be added as a Concept[3]; if it doesn't know either, both items will be added. Then it will add a link between those Concepts, with a numeric tag attached—the frequency of the 2-gram as taken from the corpus.

Over time, each Concept will end up with a set of Links to all the other Concepts it has been in an Expression with, as shown in Fig. 1. This associative structure constitutes the knowledge organization of comRAT-C. When three words are given to comRAT-C, as they would in the context of a RAT query, each of these words activates the set of Concepts they are linked to. An overlapping activation starting from two or three of the initial words can sometimes be observed. In Fig. 1, the initial words are depicted in green - COTTAGE, SWISS, and CHEESE. A two-item convergence of activation is observed for the word CHOCOLATE, and a three-item convergence for the word CHEESE. The three-item convergences are possible RAT query responses. Multiple two and three-item convergences are of course possible for the same query.

The comRAT-C system performs well in answering compound RAT queries even without the links between Concepts being weighted using the frequency of 2-gram metrics; however, adding these metrics improves comRAT-C's performance and helps break ties between multiple possible three-item convergences.

An analysis of comRAT-C's probability to find an answer based on the two-gram frequency data revealed a correlation to human performance normative data. This correlation shows that the frequency of 2-grams, on which the probability of finding the answer is based, might have an influence on the process of solving compound RAT queries. In order to keep compound queries controlled for the frequency variable, to check for other influences (like order), and to understand these influences in more detail, frequency-based probability data needs to be gathered on a large set of queries. A large enough set of queries can also be used to keep part of the query words or the answer word (different queries, same answer) constant. In order to gather such data, and construct a large set of queries, we proposed reverse-engineering our computational approach in order to generate new compound RAT items.

[3]The reason we call these items Concepts is that we entertain the possibility that they can be more than linguistic items, and another system - comRAT-V (Olteţeanu, Gautam, & Falomir, 2015)- aims to solve a visual variant of the RAT using visual associations.

Table 1 An example of preliminary results, focusing on words as answers to possible queries

w_{ans}	w_q	$fr(w_{ans}, w_q)$	$\sum_{k=1}^{m}(w_q, w_k)$	$P(w_{ans}\|w_q)$
Health	animal	95	6558	0.0145
	bone	159	4048	0.0393
	brain	25	7139	0.0035
	breast	46	10302	0.0045
	child	149	16550	0.009
	city	49	16666	0.0029

In this case the answer word is *"health"*, and only a few of its 256 links are shown

Thus, instead of using the organization structure of comRAT to provide answers, this will be used to provide queries. From this vantage point, each word w_{ans} that has at least three links, let's say to words w_a, w_b, w_c, is a potential answer to a RAT query.

Generating new remote associates test items with comRAT-G

The process of generating new Remote Associates Test items unfolds as follows. All the noun–noun high frequency 2-grams in COCA are organized in Concepts and Links in the knowledge base of comRAT-G. The selection of noun–noun 2-grams is done using the UCREL CLAWS7 Tagset[4] (tags as per this tagset are provided with the 2-grams dataset). comRAT-G uses nouns alone in this current version, unlike comRAT-C, which used more parts of speech, as described in Olteţeanu and Falomir (2015). Thus the set of most frequent 2 million 2-grams is reduced to a set of 43,908 expressions.

First, comRAT-G iterates through the words and provides preliminary results, which henceforth we shall call type 1 results, a sample of which is shown in Table 1. Type 1 results consider each word as a potential answer word. Thus in Table 1, w_{ans} stands for the answer word and w_q for a potential query word which can be used to get the answer word. Terms w_q can be further integrated in a RAT query in positions w_a, w_b or w_c. The third column represents the frequency of association between the query word w_q and the answer word w_{ans}. The fourth column represents the frequency of association between the query word and any word. The fifth column represents the probability of answer w_{ans} given specific query word w_q.[5] This is calculated as shown in Eq. 1. Query words w_q are only generated for the w_{ans}, which have at least three w_q. Applying the process to get to type 1 results yields a total of ~81500 unique w_{ans}, w_q combinations, based on 9601 unique answer words.

$$P[w_{ans} \mid w_q] = \frac{fr(w_q w_{ans})}{\sum_{k=1}^{n}(w_q, w_k)} \tag{1}$$

After type 1 results have been produced and stored, the new RAT queries are generated, using a combinatoric approach. For each answer word, the set of query words are retrieved and three-word combinations are generated. This applies the well-known combinatorics formula shown in Eq. 2, with n being the number of query items connected to a specific answer word, and k being 3.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2}$$

For each possible answer with $n < 100$ (and of course $n > 2$), all unique combination triples are produced. We capped n at 100 because of computational costs ($\binom{100}{3}$ is 161,700 possible combinations with the terms connected to the same answer word) and diminishing returns—an answer word w_{ans} connected to over 100 items might be a very common word, or form much weaker bonds with either of the words; thus RAT items constructed from its terms might not be too interesting or intuitive to solve (lower associative power of triggering result).

In order to construct all such combinations in a computationally feasible manner, comRAT-G uses Alan Tucker's combinatorics algorithm (Tucker, 2006). For 9601 query answer words and capping n at 100 (which translates into only using about 9200 answers), we obtain about 17 million possible RAT triples. The probability of answering the query is calculated based on the conditional probability of the answer to be triggered by each of the three query items. The probability thus currently considers an equal weighting of the three items, as shown in Eq. 3 (as in comRAT-C (Olteţeanu & Falomir, 2015)). However, different weighting schemes can be considered for modeling purposes—which is why we also provide the conditional probability of each

[4] http://ucrel.lancs.ac.uk/claws7tags.html

[5] Further restrictions would have been possible, for example, by only considering words as valid query words, if their conditional probability is above a certain threshold. While this may have restricted possible query words to those that have a more valid/reasonable appeal, we thought it better to start with an all-inclusive set, on which then modelers can put their own thresholds.

item, as shown in the Appendix. The various types of data items captured by this ample list of possible RAT queries and the roles in which such data can be used in empirical research are presented in "The Repository".

$$P(w_{ans}) = \frac{P(w_{ans} \mid w_a) + P(w_{ans} \mid w_b) + P(w_{ans} \mid w_c)}{3} \quad (3)$$

Evaluation

In order to check whether the queries created by comRAT-G are suitable, valid and reliable RAT queries, which can be solved by human participants and are coherent with existing RAT datasets, we have set up a study in which the human performance on comRAT-G queries is compared to that on normative data from Bowden and Jung-Beeman (2003).

Method

Two sets of query items, one comprising randomly selected comRAT-G queries and the second comprising randomly selected queries from the normative data of Bowden and Jung-Beeman (2003), were presented mixed in random order to native speakers in an online study. Accuracy and response times for solving the items were recorded. The purpose of the study was to check whether: (i) correlations between performance indicators hold between comRAT-G and Bowden & Jung-Beeman items, thus showing validity of comRAT-G items and (ii) whether comRAT-G queries are a reliable tool, as measured by Cronbach's alpha, compared to Bowden & Jung-Beeman items.

Table 2 Descriptive data on the age, education, and self-rated creativity level of the participants, $n = 113$

Indicator	Level	No. of participants
Age	Under 20	2
	20–30	33
	30–40	36
	40–50	17
	50–60	20
	60–70	4
	Over 70	1
Education level	Secondary school	7
	High school diploma	27
	Enrolled in undergraduate courses	20
	Completed undergraduate courses	34
	Enrolled in postgraduate courses	6
	Completed postgraduate courses	19
Creativity self-rated	Low	5
	Average	39
	Above average	41
	High	22
	Very high	6

Table 3 Descriptive statistics on accuracy in number of queries solved, $n = 113$

	Mean accuracy no. solved (SD)	Std. Error	95 % Conf. L.B.	interval U.B.
comRAT-G	26.20 (7.03)	0.66	24.89	27.51
Bowden & Jung-Beeman	26.41 (11.24)	1.06	24.31	28.50
Both	52.64 (16.16)	1.52	49.62	55.65

Participants

A total of 113 native English speakers, 72 female and 41 male, were recruited at University of Pittsburgh and on Crowdflower and volunteered to answer our study, which was set up online. Participants had a wide range of ages, education levels and self-rated their creativity on a wide set of levels, as shown in Table 2.

Materials

Fifty compound RAT queries were randomly selected from the items produced by comRAT-G. Another 50 queries were randomly selected from the query set of Bowden & Jung-Beeman. The comRAT-G queries can be found in the Appendix. From Bowden and Jung-Beeman (2003) we used queries 5, 6, 11, 15, 17, 20, 21, 22, 24, 26, 28, 29, 30, 37, 38, 40, 45, 46, 50, 51, 53, 58, 62, 65, 68, 71, 72, 74, 76, 79, 82, 84, 87, 90, 95, 96, 99, 106, 110, 111, 114, 116, 122, 124, 130, 131, 133, 136, 139 and 144.

Table 4 Descriptive statistics on response times (RT) in seconds for queries solved, $n = 112$

	Mean RT in seconds (SD)	Std. Error	95 % Conf. L.B.	interval U.B
comRAT-G	14.52 (9.89)	0.93	12.67	16.38
Bowden & Jung-Beeman	16.56 (12.84)	1.21	14.15	18.97

Table 5 Descriptive statistics on number of participants solving per query (of 113), and mean time spent per query (whether a correct answer was given or not)

	Mean no. of participants solving (SD)	Std. Error	95 % Conf. L.B.	interval U.B
Solved comRAT-G	59.28 (33.63)	4.76	49.72	68.84
Solved Bowden & Jung-Beeman	59.92 (21.18)	3.0	53.9	65.94
Solved both	59.6 (27.96)	2.8	54.05	65.15
	Mean time spent solving	Std. Error	95 % Conf. L.B.	interval U.B
Time comRAT-G	21.9 (10.3)	1.46	18.97	24.83
Time Bowden & Jung-Beeman	23.12 (7.1)	1.00	21.11	25.14
Time both	22.51 (8.82)	0.88	20.76	24.26

Procedure

The task was explained with two query examples. Then, five training queries were presented. These queries were taken from Bowden & Jung-Beeman items and did not overlap with our random selection of 50 items. After the participants attempted to solve the training queries, feedback including the correct answer was presented. Then, the 100 queries (50 from Bowden & Jung-Beeman, 50 from comRAT-G) were presented in random order.

Results

The dependent variables were (i) accuracy, measured as the number of correct responses for each participant in comRAT-G and Bowden & Jung-Beeman queries, and (ii) response times, measured as the number of seconds each participant spent on answering each comRAT-G and Bowden & Jung-Beeman query.

As Table 3 shows, the mean accuracy was 26.20, standard deviation (SD) = 7.03) problems correctly solved (52.4 %) for comRAT-G items and 26.41 (SD = 11.24) problems correctly solved (52.82%) for Bowden & Jung-Beeman items. The mean response times for correct solutions (n = 112[6]) was 14.52 s (SD = 9.89) for comRAT-G items and 16.56 s (SD = 12.84) for Bowden & Jung-Beeman items, as shown in Table 4.

As shown in Table 5, the mean number of participants solving each comRAT-G query was 59.28, and the mean number of participants solving each Bowden & Jung-Beeman query was 59.92. The mean time spent per comRAT-G query, independent of whether it was solved or not, was 21.9 s, while the mean time spent per each Bowden & Jung-Beeman query was 23.12.

Accuracy and response times per query for the comRAT-G dataset are shown in the Appendix.

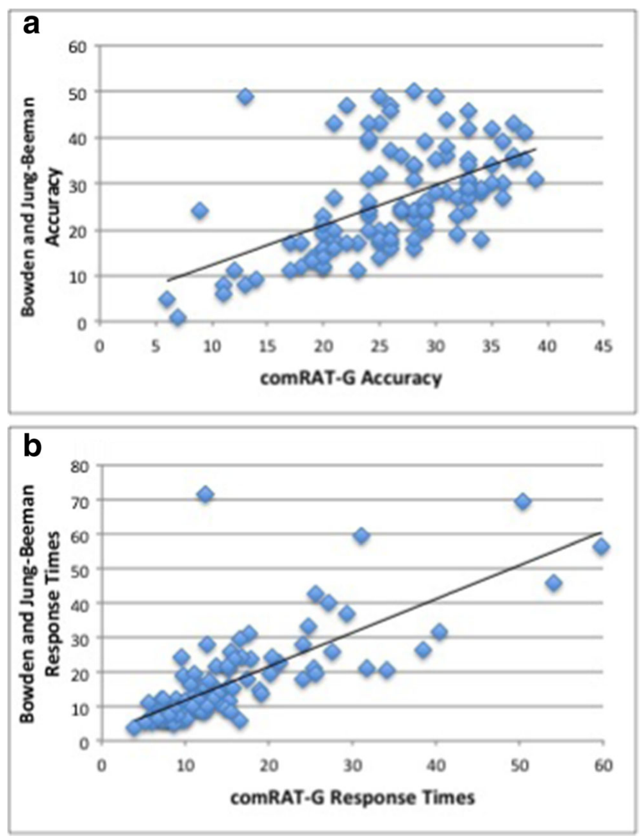

Fig. 2 Correlations on **a** accuracy and **b** response times

[6]We removed one outlier that spent 177 min on the task, as this indicated solving the RAT was not their main focus.

Accuracy showed an average significant correlation between the comRAT-G and the Bowden & Jung-Beeman datasets of $r = 0.54$, $p < 0.0001$ (Fig. 2a). Response times showed a highly significant large correlation between the two datasets of $r = 0.75$, $p < 0.0001$ (Fig. 2b). Note that response times were calculated only for correct answers.

We then measured the scale reliability of the two datasets (comRAT-G items and Bowden & Jung-Beeman items) using Cronbach's alpha as an internal consistency measure. As Table 6 shows, Cronbach's alpha on accuracy was 0.851 for comRAT-G items, 0.932 for Bowden & Jung-Beeman items and 0.936 for both sets of queries taken together. Cronbach's alpha on response times (on both correct and incorrect answers) was 0.991 for comRAT-G items, 0.99 for Bowden & Jung-Beeman items and 0.995 for both set of queries taken together.

As a final point, we checked to see whether the accuracy and response times data we obtained on the Bowden & Jung-Beeman dataset with our participants correlated with that obtained by Bowden & Jung-Beeman. As shown in Table 7, all accuracy measures and all but one response times measures correlated significantly.

Discussion

The descriptive data is similar between the comRAT-G and Bowden & Jung-Beeman's set of queries, on both mean accuracy and mean response times. The average and high correlations obtained between the performance of the participants on the comRAT-G and Bowden & Jung-Beeman sets of items on both accuracy and response times show the *validity* of the comRAT-G dataset, pointing to the fact that we are measuring the same skill with comRAT-G as with Bowden & Jung-Beeman sets. The high Cronbach's alpha internal consistency scores, which remain the same or increase when putting the two item sets together, show that both sets are highly *reliable*, and consistent with each other. Thus the comRAT-G data are in all crucial respects similar to the established query set.

Table 6 Cronbach's alpha internal consistency measures

Set of queries	Performance measure	n	Cronbach's Alpha
comRAT-G	Accuracy	113	0.851
Bowden & Jung-Beeman	Accuracy	113	0.932
Both	Accuracy	113	0.936
comRAT-G	Response times	112	0.991
Bowden & Jung-Beeman	Response times	112	0.99
Both	Response times	112	0.995

Table 7 Correlation of performance on the Bowden & Jung-Beeman queries, between our participants and Bowden & Jung-Beeman's participants

Measure type	Bowden & Jung-Beeman measure	Our measure	Correlation
Accuracy	Accuracy 2s	Accuracy	$r = 0.66$, $p < 0.0001$
	Accuracy 7s	Accuracy	$r = 0.76$, $p < 0.0001$
	Accuracy 15s	Accuracy	$r = 0.78$, $p < 0.0001$
	Accuracy 30s	Accuracy	$r = 0.81$, $p < 0.0001$
Response times	RT 7s	RT	$r = 0.45$, $p < 0.0001$
	RT 15s	RT	$r = 0.15$, –
	RT 30s	RT	$r = 0.72$, $p < 0.0001$

The Repository

In the following, the RAT queries generated and the data which accompanies them will be explained as a function of data items (columns), ability to search for and order items and some examples of possible empirical research using this data. Table 8 shows a sample of the generated queries data and their form.

The generated compound RAT queries can thus be ordered in the following ways:

(1) Alphabetically by the first, second and third word (on w_a, w_b and w_c). This ability to search for alphabetically ordered RAT queries allows empirical research keeping the first letter or the entire query word (or more than one word) constant. This can be used in various forms, at its extension allowing for keeping the entire query constant and checking for different possible answers.

(2) By the answer (w_a). This can allow for comparisons of query difficulty in which the query terms differ, and the answer is kept constant.[7] Thus for the queries a) HEALTH, CHILD and CENTER and b) INSURANCE, HAIR and CHILD, the answer is the same - CARE, and so is one of the given terms, CHILD. However, the likelihood of reaching this answer is not the same. Keeping the answer the same can help check upon the influence the different terms and their frequency have on the performance.

(3) By frequency of the favorable cases ($fr(w_a, w_{ans})$, $fr(w_b, w_{ans})$, $fr(w_c, w_{ans})$).

[7]This can also allow for keeping just the initial part of the answer constant, and checking for phonetico-syntactical rather than semantical influences in answer difference.

Table 8 An example of generated queries, organized by w_{ans} which is *ability*

w_a	w_b	w_c	$fr(w_{ans}, w_a)$	$fr(w_{ans}, w_b)$	$fr(w_{ans}, w_c)$	$fr(w_a)$	$fr(w_b)$	$fr(w_c)$	$P(w_{ans}\|w_a)$	$P(w_{ans}\|w_b)$	$P(w_{ans}\|w_c)$	$P(w_{ans})$	w_{ans}
coping	language	leadership	88	58	76	1237	6405	4556	0.0711	0.0091	0.0167	0.0323	ability
coping	language	problem-solving	88	58	50	1237	6405	606	0.0711	0.0091	0.0825	0.0542	ability
coping	language	reading	88	58	65	1237	6405	4292	0.0711	0.0091	0.0151	0.0318	ability
coping	language	reasoning	88	58	37	1237	6405	121	0.0711	0.0091	0.3058	0.1287	ability
coping	language	student	88	58	35	1237	6405	26153	0.0711	0.0091	0.0013	0.0272	ability
coping	language	thinking	88	58	34	1237	6405	1239	0.0711	0.0091	0.0274	0.0359	ability
coping	language	writing	88	58	29	1237	6405	2515	0.0711	0.0091	0.0115	0.0306	ability

(4) By the (sum) frequency of the given words ($fr(w_a)$, $fr(w_b)$, $fr(w_c)$). This allows for the study of frequency based influence separately from probability. Empirical exploration of whether keeping the frequency constant throughout the words (query or answer) among different queries has an impact on answer performance, and the function and interinfluences between frequency and answer performance should thus be possible.

(5) By the probability of the answer to be found. Multiple queries with similar probabilities can thus be analyzed, and queries from different probability classes can be analyzed (low probability, medium probability, high probability) together with their influence on human performance.

The frequency of the various query words with the answer, and the frequency of the query words occurring in other combinations has been provided here separately, as the probability of finding the answer has been calculated here taking the influence from the three words to be equal. It might be the case that the first two words have a higher influence (see Olteţeanu, 2014), and showing frequency explicitly for each of the query words allows the study of order effects.

An interface permitting access to the queries constructed by comRAT-G can be found here: http://creacogcomp.com/comRAT-G.html.

Conclusions and future work

An ample set of about 17 million queries was generated using a variant of a computational RAT solver - comRAT-G. This set of queries aims to fill a gap in the area of providing normative frequency-based data and an ampler set of stimuli for cognitive and computational creativity research. Frequency and frequency-based probability of finding the answer have been computed for all the generated queries and are provided with this data. The contributed repository allows further control over variables when testing for the influence of frequency, keeping words constant and word order in Remote Associates queries.

The entire list of queries or a subset thereof can be obtained by contacting the authors. As future work, we will aim to make the following contributions:

(i) Improve the online interface with more search and selection features, for easy access.
(ii) Generate an updated version of this repository by also parsing compound nouns from the corpus automatically, and offer queries based on compound nouns as another controllable variable. The motivation for this is that items of the form (w_{ans}, w_q) which are parsed from compound nouns might be associated more tightly than items which have 2-grams as a point of origin.
(iii) Generate a version of the repository which includes queries made of other parts of speech than nouns alone.
(iv) Offer the ability to collapse plural and singular forms to modelers.
(v) Add free association norms data to the query-answer pairs, if available.
(vi) Enable control of query and answer word length.
(vii) Enable control over semantic domain of words.
(viii) Rate a part of these queries for interestingness and hardness, in order to further refine the generating algorithm.

Acknowledgements Ana-Maria Olteţeanu gratefully acknowledges the support of the Deutsche Forschungsgemeinschaft (DFG) for the Creative Cognitive Systems (CreaCogs) project.[8] The support offered by the RISE DAAD program is acknowledged by Ana-Maria Olteţeanu and Jonathan B. Dyer. We thank Thansuda Kraisangka from Mahidol University for helping us create the interface to the generated test items.

[8] http://creacogcomp.com

Appendix: Accuracy and response times per query[9]

No.	w_a	w_b	w_c	(w_{ans}, w_a)	(w_{ans}, w_b)	(w_{ans}, w_c)	$fr(w_a)$	$fr(w_b)$	$fr(w_c)$	$P(w_{ans}\|w_a)$	$P(w_{ans}\|w_b)$	$P(w_{ans}\|w_c)$	$P(w_{ans})$	w_{ans}	Solved (n = 113)	RT (SD)
1	box	panes	shades	86	57	79	8815	123	108	0.0098	0.4634	0.7315	0.4016	window	79	13.39 (18.16)
2	bicycle	pawn	photo	45	104	35	323	149	2936	0.1393	0.6980	0.0119	0.2831	shop	94	14.1 (14.72)
3	penalty	suit	toll	5195	44	1045	6127	3016	1471	0.8479	0.0146	0.7104	0.5243	death	8	11.85 (8.74)
4	paddle	roulette	steering	27	57	1809	79	95	2469	0.3418	0.6000	0.7327	0.5582	wheel	91	12.1 (10.98)
5	gala	invitation	table	55	29	1120	130	90	16357	0.4231	0.3222	0.0685	0.2713	dinner	15	27.51 (29.17)
6	cellars	regions	shop	29	27	61	55	231	7480	0.5273	0.1169	0.0082	0.2175	wine	77	18.24 (26.99)
7	closing	departure	lunch	84	40	75	195	107	2903	0.4308	0.3738	0.0258	0.2768	date	35	14.92 (14.45)
8	cedar	fig	bark	40	122	126	172	351	342	0.2326	0.3476	0.3684	0.3162	tree	95	11.76 (13.98)
9	musicians	protests	shoes	42	130	30	418	195	1814	0.1005	0.6667	0.0165	0.2612	street	16	50.71 (46.92)
10	gain	prosperity	wealth	50	38	120	1232	77	797	0.0406	0.4935	0.1506	0.2282	material	1	15.83 (–)
11	ocean	tile	level	504	327	72	2075	634	15700	0.2429	0.5158	0.0046	0.2544	floor	52	15.89 (17.76)
12	pep	pillow	host	263	70	26	558	203	3466	0.4713	0.3448	0.0075	0.2745	talk	75	13.8 (19.85)
13	bait	hatchery	tanks	27	37	54	89	77	1322	0.3034	0.4805	0.0408	0.2749	fish	92	10.36 (12.23)
14	pounding	attacks	defect	62	1204	54	710	3394	167	0.0873	0.3547	0.3234	0.2551	heart	77	13.57 (12.13)
15	harmony	saturation	schemes	36	33	107	75	263	274	0.4800	0.1255	0.3905	0.3320	color	26	20.41 (19.6)
16	enemies	scoring	word	56	25	43	89	606	2799	0.6292	0.0413	0.0154	0.2286	list	2	26.67 (25.71)
17	dealers	industry	mechanic	104	1232	167	2128	28295	212	0.0489	0.0435	0.7877	0.2934	auto	100	8.57 (7.54)
18	complaints	fraud	tastes	136	52	36	414	1845	73	0.3285	0.0282	0.4932	0.2833	consumer	15	64.58 (105.22)
19	clippings	columns	story	235	68	129	557	244	9643	0.4219	0.2787	0.0134	0.2380	newspaper	96	11.83 (14.54)
20	cartoon	actors	flaws	197	69	68	565	307	136	0.3487	0.2248	0.5000	0.3578	character	66	14.31 (11.57)
21	banker	plan	strategies	640	53	63	672	14504	3159	0.9524	0.0037	0.0199	0.3253	investment	29	15.08 (9.8)
22	cup	crate	powder	415	62	55	13087	93	3222	0.0317	0.6667	0.0171	0.2385	milk	5	24.45 (17.42)
23	checking	escrow	deficits	337	110	34	405	134	976	0.8321	0.8209	0.0348	0.5626	account	82	15.58 (18.48)
24	holiday	window	mall	251	42	867	2469	7884	1306	0.1017	0.0053	0.6639	0.2570	shopping	75	16.37 (12.97)
25	bikes	climber	peak	328	73	63	479	184	809	0.6848	0.3967	0.0779	0.3865	mountain	100	10.41 (8.68)
26	climbers	formation	song	67	96	56	129	1437	2305	0.5194	0.0668	0.0243	0.2035	rock	40	19.16 (26.71)
27	health	track	labels	35	1609	249	79202	4132	741	0.0004	0.3894	0.3360	0.2419	record	6	19.1 (10.33)
28	baseball	neighborhood	ranger	45	59	217	9730	2059	369	0.0046	0.0287	0.5881	0.2071	park	47	21.43 (26.86)
29	cold	dancing	hockey	88	41	341	222	465	2406	0.3964	0.0882	0.1417	0.2088	ice	89	18.09 (53.18)/ 12.72 (17.17)
30	airplane	solo	simulator	33	33	106	468	1460	135	0.0705	0.0226	0.7852	0.2928	flight	75	15.48 (16.54)
31	clip	footage	games	306	128	1820	797	825	8054	0.3839	0.1552	0.2260	0.2550	video	41	14.17 (11.7)
32	creature	foam	snakes	70	33	47	212	491	119	0.3302	0.0672	0.3950	0.2641	sea	57	18.84 (29.65)/ 15.49 (15.98)
33	embargo	fire	shipments	572	53	75	1049	9515	213	0.5453	0.0056	0.3521	0.3010	arms	6	26.05 (19.41)
34	florist	hangers	whisk	30	28	136	60	88	160	0.5000	0.3182	0.8500	0.5561	wire	24	47.8 (62.79)/ 39. 97 (50.82)
35	departure	turning	guard	43	1090	1808	107	1395	4789	0.4019	0.7814	0.3775	0.5203	point	18	25.17 (28.43)
36	breeze	urchins	voyage	97	173	53	210	202	108	0.4619	0.8564	0.4907	0.6030	sea	92	9.8 (13.83)
37	chores	incomes	pets	302	132	51	351	252	113	0.8604	0.5238	0.4513	0.6118	household	76	23.58 (23.5)
38	blades	coaster	skates	29	1026	97	1113	1125	155	0.0261	0.9120	0.6258	0.5213	roller	98	8.56 (8.15)
39	blouse	handkerchief	ribbon	162	55	26	230	84	155	0.7043	0.6548	0.1677	0.5089	silk	43	14.85 (11.05)
40	math	obstacle	refresher	30	340	95	2484	389	124	0.0121	0.8740	0.7661	0.5507	course	81	12.39 (12.84)
41	lot	spaces	ticket	8264	288	112	8832	472	3641	0.9357	0.6102	0.0308	0.5256	parking	94	9.34 (9.21)
42	juice	tart	zest	3506	44	601	9568	107	758	0.3664	0.4112	0.7929	0.5235	lemon	70	8.95 (6.14)
43	deterrent	ripple	snowball	110	320	40	135	434	116	0.8148	0.7373	0.3448	0.6323	effect	59	24.46 (41.53)/ 20.46 (28.45)
44	brown	cookies	mousse	69	37	203	97	552	235	0.7113	0.0670	0.8638	0.5474	chocolate	99	16.57 (20.7)
45	nest	wash	yolk	382	64	346	657	553	416	0.5814	0.1157	0.8317	0.5096	egg	99	9.85 (7.76)
46	bolts	bugs	storm	131	79	89	156	104	2802	0.8397	0.7596	0.0318	0.5437	lightning	61	29.38 (104.9)/ 16.08 (14.72)
47	nucleus	phones	tumors	31	2109	31	57	2466	295	0.5439	0.8552	0.1051	0.5014	cell	75	19.93 (20.7)
48	concern	discrepancy	lifting	33	38	197	268	64	226	0.1231	0.5938	0.8717	0.5295	weight	18	37.1 (57.56)/ 17.61 (10.47)
49	apnea	deprivation	pattern	332	197	30	361	308	1045	0.9197	0.6396	0.0287	0.5293	sleep	98	6.89 (4.9)
50	jump	resorts	slope	71	165	81	790	189	122	0.0899	0.8730	0.6639	0.5423	ski	95	7.37 (7.12)

[9] Response times and SD for queries 29, 32, 34, 43, 46 and 48 have also been shown with corrections for outliers. All the data analyses have been done without these corrections.

References

Ansburg, P. I., & Hill, K. (2003). Creative and analytic thinkers differ in their use of attentional resources. *Personality and Individual Differences*, *34*(7), 1141–1152.

Arden, R., Chavez, R. S., Grazioplene, R., & Jung, R. E. (2010). Neuroimaging creativity: a psychometric view. *Behavioural Brain Research*, *214*(2), 143–156.

Baba, Y. (1982). An analysis of creativity by means of the remote associates test for adult revised in Japanese (JARAT FORM a). *The Japanese Journal of Psychology*, *52*(6), 330–336.

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

Bringsjord, S. (2011). Psychometric artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, *23*(3), 271–277.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences*, *106*(25), 10130–10134.

Chermahini, S. A., 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*(3), 177–186.

Duncker, K. (1945). On problem solving. *Psychological Monographs*, *58*(5), i–113.

Fodor, E. (1999). Subclinical inclination toward manic-depression and creative performance on the remote associates test. *Personality and Individual Differences*, *27*(6), 1273–1283.

Guilford, J. P. (1967). The nature of human intelligence. McGraw-Hill.

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance tests of creative thinking (TCT). *Creativity Research Journal*, *18*(1), 3–14.

Maier, N. R. (1931). Reasoning in humans. ii. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, *12*(2), 181.

Mednick, S. A., & Mednick, M. (1971). Remote associates test: Examiner's manual. Houghton Mifflin.

Olteţeanu, A.-M. (2014). Two general classes in creative problem-solving? An account based on the cognitive processes involved in the problem structure - representation structure relationship. In Besold, T., Kühnberger, K. U., Schorlemmer, M., & Smaill, A. (Eds.) *Proceedings of the international conference on computational creativity*, Vol. 01-2014: Osnabrück.

Olteţeanu, A.-M. (2016). From simple machines to eureka in four not-so-easy steps. Towards creative visuospatial intelligence. In Müller, V. (Ed.) *Fundamental issues of artificial intelligence*. https://doi.org/10.1007/978-3-319-26485-1_11, (Vol. 376 pp. 159–180): Springer.

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., & Falomir, Z. (2016). Object replacement and object composition in a creative cognitive system. Towards a computational solver of the alternative uses test. *Cognitive Systems Research*, *39*, 15–32. https://doi.org/10.1016/j.cogsys.2015.12.011.

Olteţeanu, A.-M., & Schultheis, H. (in press). What determines creative association? revealing two factors which separately influence the creative process when solving the remote associates test. *The Journal of Creative Behaviour*. https://doi.org/10.1002/jocb.177.

Olteţeanu, A.-M., Gautam, B., & Falomir, Z. (2015). Towards a visual remote associates test and its computational solver. In *Proceedings of the third international workshop on artificial intelligence and cognition 2015*, (Vol. 1510 pp. 19–28): CEUR-Ws.

Olteţeanu, A.-M., Falomir, Z., & Freksa, C. (in press). *Artificial cognitive systems that can answer human creativity tests: an approach and two case studies. IEEE Transactions on Cognitive and Developmental Systems*. https://doi.org/10.1109/TCDS.2016.2629622.

Qiu, J., Li, H., Yang, D., Luo, Y., Li, Y., Wu, Z., & Zhang, Q. (2008). The neural basis of insight problem solving: an event-related potential study. *Brain and Cognition*, *68*(1), 100–106.

Saugstad, P., & Raaheim, K. (1957). Problem-solving and availability of functions. *Acta Psychologica*, *13*, 263–278.

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

Sio, U. N., & Rudowicz, E. (2007). The role of an incubation period in creative problem solving. *Creativity Research Journal*, *19*(2-3), 307–318.

Sitton, S. C., & Pierce, E. R. (2004). Synesthesia, creativity and puns 1. *Psychological Reports*, *95*(2), 577–580.

Storm, B. C., Angello, G., & Bjork, E. L. (2011). Thinking can cause forgetting: memory dynamics in creative problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1287.

Tucker, A. (2006). *Applied combinatorics*. Wiley.

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

Ward, J., Thompson-Lake, D., Ely, R., & Kaminski, F. (2008). Synaesthesia, creativity and art: What is the link? *British Journal of Psychology*, *99*(1), 127–141.

Whitt, J. K., & Prentice, N. M. (1977). Cognitive processes in the development of children's enjoyment and comprehension of joking riddles. *Developmental Psychology*, *13*(2), 129.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & Cognition*, *26*(4), 716–730.

Worthen, B. R., & Clark, P. M. (1971). Toward an improved measure of remote associational ability. *Journal of Educational Measurement*, *8*(2), 113–123.