Computationally resurrecting the functional Remote Associates Test using cognitive word associates and principles from a computational solver

Ana-Maria Olteţeanu

Cognitive Systems, Human-Centered Computing, Freie Universität Berlin, Germany

Mikkel Schöttner

Freie Universität Berlin, Germany

Susanne Schuberth

Universität Osnabrück, Germany

Abstract

Human creativity is usually assessed with a variety of established creativity tests. One of this is the Remote Associates Test (RAT), which aims to measure the ability of reaching remote associates with linguistic stimuli. A well known variant of the RAT exists – the compound RAT, for which normative data and solvers have been proposed in the literature. However, a different type of RAT was proposed in 1971 by Worthen and Clark – a functional form which had the potential of measuring other types of associations. However, the few test items proposed by Worthen and Clark where lost during archive transport, and cannot be accessed. In this paper, we set to reconstruct an ample set of functional items in the spirit of Worthen and Clark's idea, using information science techniques. Cognitive word associates are used as data. The process of a former computational solver of the RAT is repurposed to create rather than solve items. The approach of constructing queries is evaluated by getting human participants to solve both functional and compound items. In the process, a

---

*Corresponding author
Email address:* `ana-maria.olteteanu@fu-berlin.de` (Ana-Maria Olteţeanu)

previous computational approach to solving the compound RAT is also validated in the functional RAT context.

1. Introduction

Human creativity [1, 2] is a complex set of cognitive processes, assessed in a multitude of ways, with different creativity tests measuring its various and sometimes overlapping aspects [3, 4, 5, 6, 7, 8, 9]. The Remote Associates Test [10] – RAT for short – aims to measure a participant's creativity as their ability to make associations. The RAT is normally given using linguistic stimuli, though visual stimuli have also been developed [11]. A Remote Associates Test item is administered as follows: three words are given, and the participant is asked to come up with a fourth word which is related to all the three given words. For example, the words OPERA, HAND and DISH are given. To this RAT item, the word SOAP can be an answer, because of the existing relationships between the answer word and the query words, which can be observed in the linguistic terms SOAP OPERA, HAND SOAP and DISH SOAP.

The Remote Associates Test is widely used [12, 13, 14, 15] and has been constructed and normed in a variety of languages, including: English [16]; Italian [17]; Dutch [18]; German [19]; Japanese [20], Chinese [21] and Polish [22].

The Remote Associates Test is based on Mednick's remote association theory [23]. While in accord with the basic principles of this theory, Worthen and Clark [24] have argued that the items generated for the Remote Associates Test could be further fine-tuned; on the one hand, they claimed that the test consisted of at least two types of queries. On the other hand, they remarked on the fact that not all the associates used were part of the Palermo-Jenkins word association norms [25].

The two types of items Worthen and Clark distinguished were (a) language-based associations (*structural*), and (b) associations that stem from connections

2

beyond language (*functional*). For instance, a structural association is STREET and FOOD. This association could be incorporated in a structural RAT item. Contrarily, items like BIRD and FEATHER yield a functional relationship. Hence, the query OPERA, HAND and DISH is a structural query, as each word is lin-
guistically connected to the word SOAP. A functional query would, in turn, be based on functional relationships, whether or not these words occur as linguistic patterns. For example, functional relationships like "bear-animal": is-a relationship (hyponym-hypernym), "screen-monitor": is-a-part-of relationship (meronym-holonym), "smell-scent": (synonyms) may be part of such functional
queries.

However, the normative data on the Remote Associates test refers to items which are linguistic associates [16]; these items are called *compound* by Bowden and Jung-Beeman, because the query words form compounds with the answer. It is worth noting that compound words are a subcategory of potential language
based association items - with Worthen and Clark's definition. Other linguistically occurring patterns can also be, as one of our reviewers points out, adjacent words, or words occurring in the same phrase.

Furthermore, computational efforts at both solving [26] and generating the Remote Associates Test [27] have so far been centered around structural items
in the Worthen and Clark sense, of linguistically related terms, including compound items as a subset. The computational solver comRAT-C [26] for instance extracts its knowledge from language corpuses and generates good answers to *compound* queries that correlate to human performance. For an overview of these and other computational models of creativity like CreaCogs[28, 29],
MicroPsi/Psi[30] and IDyOT[31] see [32, 33].

Worthen and Clark transformed 20 of Mednick's queries to functional items. Yet, this set of items was impossible to retrieve and is most likely lost. The annex of Worthen and Clark's paper, containing the functional associates, was stored as part of the former National Auxiliary Publications Service (NAPS); after its
dissolution, the items were transferred to the Library of Congress. However, according to the Library of Congress, the item collection never arrived there

3

(personal correspondence 21st of July 2016). Contacting the authors'universities has not yet led to any further clues about the preservation of the queries. Hence currently, there are no datasets of *functional* queries or the respective normative data available. Worthen and Clark [24] further emphasized the relevance of the Palermo-Jenkins word association norms for the RAT queries, making the point that the absence of norms for a particular test item might indicate that the item is not suitable.

These issues have not yet been addressed in the creative problem solving literature. The existence of a functional set of RAT items would allow Worthen and Clark's theoretical work to be continued, enabling researchers to compare human performance on the different types of queries, to study whether the process of solving the two types is indeed different, and help further refine theories on creative problem solving processes.

In this paper, the authors use the two points put forward by Worthen and Clark creatively, setting to computationally construct a set of functional RAT items using a modern set of word association norms [34]. The article first describes preliminary work in Section 2 by illuminating three points: (i) the comRAT-C approach [26] which is the base of the functional RAT, (ii) the distinction of functional and compound items, and (iii) the relevance of word association norms [34]. The procedure of generating queries based on word associates is explained in Section 3. Section 4 presents the results obtained after applying this procedure. These outcomes are evaluated with human participants, in comparison to the participants' performance in the compound Remote Associates Test (Section 5). Section 6 discusses the limits of the approach and possible future extensions.

2. Preliminary work

The work on computationally constructing functional Remote Associates Test items is related to the first author's previous work on a computational solver of the compound RAT – comRAT-C [26].

4

The comRAT-C cognitive system solves compound RAT queries by extracting linguistic knowledge and organizing this knowledge in a specific associative manner. As linguistic knowledge, n-grams from a language corpus – the Corpus of Contemporary American English (COCA): `http://corpus.byu.edu/coca/` – are added as expressions to comRAT-C's knowledge base. In comRAT-C's knowledge organization, expressions are considered to be links between two concepts. These links are derived from existing expressions or compound words, where two words appeared together. For example, in Figure 1, a link exists between words SWISS and ALPS because they have co-occured in the set of n-grams comRAT-C has been exposed to. Such links between words build the foundation of the network comRAT-C generates.

After building this knowledge base network, comRAT-C uses convergence over its associative knowledge to solve RAT queries. An example of this solving process is visually depicted in Figure 1: The green items (COTTAGE, SWISS and CAKE) are the query to be solved and serve as seed concepts. To find the target word, comRAT-C activates the neighborhood associated nodes (depicted in blue) of each of the seed concepts. Thus CHOCOLATE is activated by query words SWISS and CAKE, while CHEESE is activated by all three. The first found 3-item convergence wins – thus in this case, comRAT-C would propose CHEESE as a potential answer. If no 3-item convergence is found, 2-item convergences can be returned as answers. The comRAT-C cognitive system could thus propose CHOCOLATE in the case depicted in Figure 1 as a potential answer.

The links between concepts can also be weighted using the frequency of the linguistic expressions or compounds. Using weights, the likelihood of specific answers can be computed, and the comRAT-C system can provide multiple answers to compound queries [35]. The performance of the system correlates to that of humans, specifically to the accuracy and response times in the normative dataset of Bowden and Jung-Beeman [16].

Instead of computationally solving the RAT, like the comRAT-C system, this paper focuses on generating functional RAT queries computationally. In previous work, the manual construction of visual RAT queries has been attempted,
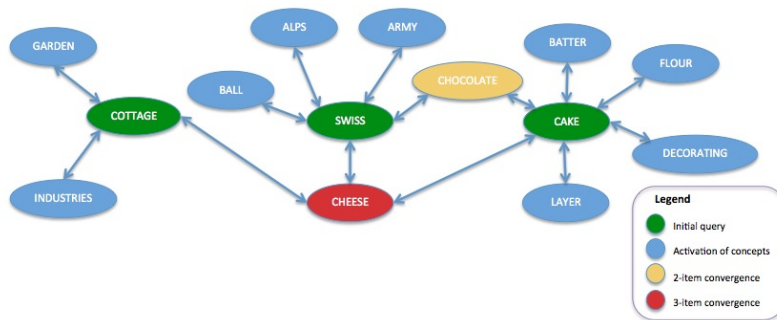
5

Figure 1: Visualization of the comRAT-C knowledge organization (with limited links for clarity), and of activation during a query.

to provide a cross-modal perspective on the RAT [11]. In terms of linguistic queries, computational query generation has been shown to be both successful [27] and useful for the design of experiments with a higher degree of control [36]. However, functional RAT queries have never been constructed previously, as only Worthen and Clark's suggestion exists, but no initial dataset has been preserved.

In order to build a *functional* form of the RAT, this approach uses a source of functional, rather than linguistic connections (n-grams). Following Worthen and Clark [24] who suggested using the Palermo-Jenkins word association norms to evauate the RAT items, we decided to extract the required knowledge for building functional RAT items using another collection of association norms by Nelson et al. [34]. This dataset contains norms for free association, word fragments, and rhyme. The dataset records how many participants presented with a cue word produce a target word. For example, when presented with the cue word ABUNDANCE, 2 of the 152 participants mentioned the word FAMINE; 5 participants mentioned the word FOOD; 3 participants mentioned the word FULL.

As it is unlikely that expressions like *abundance famine* and *abundance food* would often occur as linguistic compounds, but some form of non-linguistic relationship exists between these associates, we have considered them a good

6

source of data for functional relationships.

3. Generating queries based on word associates

The preliminary work on comRAT-C[26, 32] has been used to provide con-
ceptual insight about knowledge organization useful to the construction of func-
tional queries. The comRAT-C system solves the RAT with ease because of its
knowledge organization: if two concepts co-occured in an expression or word,
they are linked to each other and can easily be activated. This knowledge orga-
nization can not only be used to find 3-word convergences, it is also applicable
for the generation of new queries: all words with more than three links are
potential answers to a RAT query. Figure 2 shows for example that the word
SWISS is a possible answer for the query ALPS, CHOCOLATE, CHEESE. Note
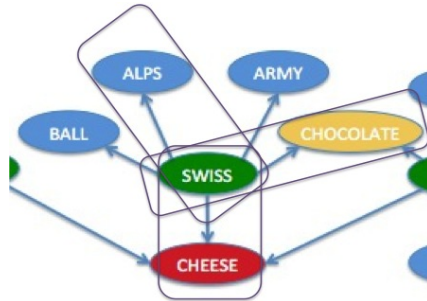that SWISS previously served as a query word.

Figure 2: Shifting from query solving to query generation: The former query word SWISS
becomes the answers to a new triple of items that is derived from the links existing in the
comRAT-C knowledge structure.

The approach for generating functional RAT queries proposed here uses this
conceptual insight on knowledge organization and the word associates dataset
to provide functional relationships as follows. In two steps, functional RAT
queries were created using word associates:

Step 1 – Items from the word associates dataset are extracted and the
knowledge is organized. An answer-centered view is applied, which focuses on
word items as answers to potential query items. Thus all items that have more

7

than three associates are considered potential answers. For future reuse, all potential answers are saved in a database table, together with their potential query items, as shown in Table 1. For each such item stand-ins for frequency and probability metrics from the University of South Florida association norms are added to the table (explained below).

A Note on stand-ins for frequency and probability. The comRAT-C [26] cognitive system transformed 2-grams like SWISS ALPS in two linked concept nodes of the expression. The link was tagged with the number of times the expression appeared in the corpus. To determine the probability that an answer for a certain item would be generated, the ratio of responses appearing in conjunction with an item over the total instances of the query item occurrences was computed.

In the context of producing a set of functional RAT items, some of the data provided by the University of South Florida association norms could be used as a substitute for frequency and probability. Thus, (i) the number of times the target is produced by the subjects can be used in lieu of frequency; and (ii) the forward strength (number of participants producing the target in response to a cue divided by the number of times participants were given the cue) can be interpreted as the probability, given that it expresses the same relationship between favourable times over total times.

Table 1: Example of answer centered view *Abundance*.

Answer	Query item	Sample size	Subjects producing target	Forward strength
Abundance	Famine	152	2	0.013
Abundance	Food	152	5	0.033
Abundance	Full	152	3	0.02
Abundance	Lack	152	4	0.026
Abundance	Large	152	2	0.013
Abundance	Little	152	2	0.013
Abundance	Lots	152	34	0.224
Abundance	Many	152	12	0.079
Abundance	Money	152	6	0.039

Step 2 – For each potential answer, all possible three-word combinations

8

yielding the answer are produced using a combinatorics algorithm from Alan Tucker [37].

The system was named $comRAT - G_F$, expressing both the conceptual relationship to the solver for the compound RAT (comRAT-C) and the distinct focus on creating functional queries.

4. Results - functional Remote Associates Test items

A number of concept pairs equal to the University of Florida free associations dataset was collected (72186 items). After that, potential queries and their answers were gathered, and items with less than three associates were excluded; a set of 72186 items remained and yielded 13 534 865 potential items for the functional RAT. The results will be analyzed in this section.

Examples of queries generated with $comRAT - G_F$ can be found in table 2. Most of them are functional queries in the sense of Worthen and Clark: the connection between query items and answers goes beyond language. A closer look shows that some of the queries are a mix of functional relationships and relationships that are both functional and linguistic, e.g. in Q4 with the query words DAISY, TULIP, VASE that have the answer FLOWER. The pairs (DAISY, FLOWER) and (TULIP, FLOWER) yield a functional relationship in the sense that the answer defines the category of the item, incorporating an *is a* relationship. However, the pair (VASE, FLOWER) happens to appear both in language, namely in the compound FLOWER VASE, and in the physical experience, as flowers go in vases.

This indicates that queries from word associates lists are not necessarily all and purely functional, thus results do contain some queries that are (also) compound queries, for example BANK ACCOUNT, BANK TELLER, BANK VAULT.

Another type of relation found is synonymy. As the query ATTENDANCE, CONTEMPORARY, GIFT with the answer PRESENT examplifies, these synonyms can span over different semantic domains, here *being present at an event*, *being of the present time* and *a present for someone*. Thus the query reveals versions

9

of the homonym PRESENT, which are semantically distinct from each other.

Table 2: A subset of functional Remote Associates items generated with $comRAT - G_F$. w_{ans} is the answer word, while $w_1 - w_3$ are the query items. Items are ordered based on their probabilities.

Query no.	w_1	w_2	w_3	w_{ans}	Probability
1	exhausted	sleepy	weary	tired	0.7202
2	frame	photo	portrait	picture	0.6897
3	bassinet	crib	infant	baby	0.6916
4	daisy	tulip	vase	flower	0.6914
5	bulb	dark	dim	light	0.5530
6	account	teller	vault	bank	0.4301
7	cashew	rat	squirrel	nut	0.3518
8	comet	limit	velocity	speed	0.2301
9	attendance	contemporary	gift	present	0.2301
10	capability	function	leadership	ability	0.1101
11	plenty	quantity	site	lot	0.0701
12	car	piston	steam	engine	0.0701
13	A	rate	test	grade	0.0501
14	agent	deception	FBI	spy	0.0163
15	earthquake	war	weakness	fear	0.0114
16	admire	jewel	ocean	beautiful	0.0111
17	cougar	go	learn	fast	0.0110
18	burn	flash	pants	down	0.0110
19	exam	flee	warn	fear	0.0110
20	condition	croak	doctor	dead	0.0105
21	case	fact	threshold	point	0.0103

$comRAT - G_F$ can create multiple queries which map to the same answer, e.g. Q15 and Q19. Note that the probabilities, which are a function of forward strength, differ across queries. Each of the three query items is weighted equally for computing this probability, but different assumptions can also be modelled. Further investigation can indicate whether the different probabilities of the queries relate to or influence human performance.

A wide spread of probabilities and frequencies is present in the created query set. The probability (mean of the forward strength over the 3 items) ranged from 0.0101, e.g. for the query BEAD, IRON, STYLE (answer: DRESS) to 0.8223 for the query SALMON, TROUT, TUNA (answer: FISH). Table 3 provides information about the distribution of items by probability.

10

Table 3: Probability distribution of the generated functional RAT items.

Probability	< 0.1	0.1 − 0.2	0.2 − 0.3	0.3 − 0.4	0.4 − 0.5	0.5 − 0.6	0.6 − 0.7	0.7 − 0.8	>= 0.8
No.	10 800 063	1 884 946	668 914	142 324	31 711	5 826	975	105	1

The frequency indicates the amount of people who returned the target word when the cue was given. The least frequent *query word − answer word* pairs were for example DECEPTION-SPY, BEAD-DRESS and LEAN-STRONG with only two productions. The most frequent associates to a specific cue were given around 134-165 times, for example: EXHAUSTED-TIRED (136), ROW-BOAT (136), HUSBAND-WIFE (146), ATTEMPT-TRY (146), 163 EAST-WEST (163) and WEEP-CRY (165). A numerical discription of the frequency distribution can be found in Table 4.

Table 4: Frequency distribution of the generated functional RAT items.

Frequency	< 10	10 − 30	30 − 50	50 − 70	70 − 90	90 − 110	110 − 130	130 − 150	>= 150
$fr(w_1, w_{ans})$	10 187 792	2 253 943	530 668	235 974	170 569	108 480	37 277	9 729	433
$fr(w_2, w_{ans})$	10 167 757	2 289 819	524 780	238 535	156 880	99 398	42 443	13 736	1517
$fr(w_3, w_{ans})$	10 150 287	2 281 637	540 365	258 690	164 917	83 566	41 647	11 613	2143

5. Evaluation with human participants

In order to evaluate the performance of humans in the generated queries and to find potential relationships in the performance between the functional and compound queries, two studies were conducted. First, an exploratory study was completed by 26 subjects that had previously solved the compound RAT, which were asked to now solve the functional RAT. This study revealed a correlation between human performance and computationally produced query probability. In order to validate the results and to gain more power, a second study was set up, in which new participants solved both compound and functional queries.

5.1. Study I

5.1.1. Method

A set of 75 items was selected from the pool of generated queries. 25 items each were derived from three batches, representing different probabilities (p) to obtain the target word (batch 1: $0.3 \le p \le 0.5$, batch 2: $0.1 \le p \le 0.3$, batch 3: $p \le 0.1$). The items were selected according to the following criteria: a) the primary relationship between words was functional; b) there were no doublings in the target words, and c) the items had to be valid across different native English speaking nationalities, hence queries that reference to local TV shows and suchlike were excluded. The queries were presented in a randomized order and time spent on each query was recorded.

5.1.2. Procedure

The participants were invited to the study via the Figure Eight crowdsourcing platform (formerly known as CrowdFlower). They were promised a monetary compensation for their efforts. After a welcome landing page, participants filled in a short questionnaire about demographics and their self-rated creativity and problem solving abilities, and gave consent for their anonymised data to be used for scientific purposes. The subjects were introduced to the task with two easy examples. They then completed five training queries, which additionally clarified how the queries are supposed to be solved; in each of these training queries it was emphasized that the target word must relate to all three given cues. The cue words were presented simultaneously and next to each other; the participants were required to type their answer in a text box below the cues.

5.1.3. Participants

All participants were advanced CrowdFlower users and had taken part in the testing of compound items computationally created with comRAT-G[27] about half a year before. 26 persons (22 females, 4 males) took part in the first study. Age was recorded using age brackets of 10 years. Three participants (11.5 %) were 20-30 years old, 14 (53.8%) were 30-40, three (11.5%) were 40-50

and six (23.1%) were 50-60 years old. Of the participants, three (11.5%) finished
secondary school and seven (26.9%) high school. Six (23.1%) completed under-
graduate courses and three (11.5%) postgraduate courses. Currently enrolled in
undergraduate courses were four (15.4%) of the participants and three (11.5%)
were enrolled in postgraduate courses. Participants self-rated their creativity
and problem solving on a 5-point Likert-scale. Their creativity ratings averaged
at 2.54 ($SD = 1.12$) and problem solving ratings at 2.81 ($SD = 0.79$).

5.1.4. Data analysis

The responses were ranked as correct (matching the target word in the gen-
erated queries) or incorrect; however, in some cases the provided answer was
semantically very closely related to the target word and had also a functional
relationship to all cues. In those cases the answer was rated as correct. For
example, the query AROMA, FUMES, GARLIC had the generated target word
SMELL. But the given answers ODOUR and SCENT yield a similarly functional
relationship to the three given words and must therefore be considered as cor-
rect. The synonymy relationship between words was rated manually by a hu-
man rater which consulted an online dictionary to verify for these relationships
- `https://www.thesaurus.com`. An computational approch may in the future
be applied, by using WordNet *synsets*; such approaches have been previously
employed successfully [38].

The number of correct answers was scored for each participant, and their
performance in the functional queries correlated to their previous performance
in solving compound queries.

5.1.5. Results

Tables 5 and 6 show the descriptive statistics on accuracy, measured as
the number of queries answered correctly and response time for correct queries.
Included are mean, standard deviation, percentage of queries answered correctly.

The scores on the fRAT items in terms of accuracy show a large positive
significant correlation with the accuracy of scores the participants obtained when

Table 5: Descriptive statistics on accuracy

RAT type - produced with	n	mean no. of queries solved (SD), percentage	Std. Error	95% CI LB	95% CI UB
functional RAT - comRAT-GF	26	35.27 (7.99), 47.03%	1.57	32.20	38.34
compound RAT - comRAT-G	46	25.02 (7.26), 50.05%	1.07	22.93	27.12

Table 6: Descriptive statistics on response times in seconds

RAT type - produced with	n	mean RT (sec) of queries solved (SD)	Std. Error	95% CI LB	95% CI UB
functional RAT - comRAT-GF	26	13.91 (8.42)	1.70	10.68	17.15
compound RAT - comRAT-G	46	12.38(6.23)	0.94	10.54	14.21

solving compound items produced with comRAT-G ($r = .55, p < .005$). A significant correlation between the performance in answering functional RAT items and compound RAT items was also observed for response times ($r = .41, p < .05$).

5.2. Study II

5.2.1. Method

As only a small number of participants could be recruited for the functional RAT study from the participants of the previous compound RAT study, a second study was conducted to further investigate possible relationships between functional and compound queries and to evaluate the performance of human participants on these.

For this study, 96 items consisting of three words each were used as stimuli. The participants were asked to provide a word that relates to all three of these. Again, only native English speakers were recruited as participants. The 96 items consisted of:

– 48 fRAT queries,

– 24 comRAT-G queries [27],

– and 24 items by Bowden and Jung-Beeman [16].

This way, an equal amount of functional items and compound items was given. The 48 functional items were chosen so that they do not contain any

14

compound words. For the compound items, strong functional relationships between the words were avoided. This was done in order for the functional and compound items to be as distinctive as possible. Additionally, there were no doublings in the target words and items with no cultural references were chosen, so that they are valid across nationalities.

A power analysis was performed a priori with G*Power. This showed that in order to reliably capture an effect size of .41 (which was the lower of the two correlations that were obtained in study I), a power of .95 and an alpha of .5, at least 55 participants would need to be recruited.

5.2.2. Procedure

The procedure was very similar to study I, with just a few differences. Firstly, participants had to solve only four training questions, two of which showed a functional, and two a compound relationship between the cues and the target words. Participants were told that they had to solve at least one test question right for the rest of their data to be considered.

Secondly, two different measures of verbal fluency were collected, using letters and categories as stimuli. In the letter paradigm, participants were asked to write down as many words starting with a given letter as they could think of in one minute. The letters used were F, A and S. For the category fluency measure, participants were asked to name as many examples in a given category as they could think of in one minute. Those categories were fruits, animals and furniture items. These tasks were given before the training questions for the RAT items.

5.2.3. Participants

A sample of 63 participants was recruited for the study, two of which had to be excluded due to not being able to answer at least one of the test questions right, thus not showing a sufficient understanding of the given task. This way, a total sample size of 61 (44 females and 17 males) was acquired, which exceeds the minimum of participants needed as indicated by the power analysis

(see above). Participants were asked to state their age on an ordinal scale with age brackets of ten years. Of the participants one (2%) was under 20 years old, 14 (23%) were 20-30, 13 (21%) were 30-40, 12 (20%) were 40-50, 15 (25%) were 50-60 and six (10%) were 60-70 years old. The majority of participants finished higher education with six (10%) having completed postgraduate and 32 (52%) undergraduate courses. One person (2%) was enrolled in postgraduate courses and four (7%) in undergraduate courses. Three participants (5%) finished secondary school and 15 (25%) obtained their high school diploma.

Participants were also asked to self-rate their creativity and problem solving skills on a 5-point Likert-scale. The mean rating for creativity was 2.92 ($SD = 1.08$), and for problem solving 2.2 ($SD = 0.95$).

5.2.4. Data analysis

In contrast to study I, only answers that matched the target word were considered correct. This was done in order to check if the results of study I would hold without considering synonyms. Answer words were matched computationally to expected answers using R.

5.2.5. Results

Tables 7 to 10 show the descriptive data of both accuracy, measured as the number of items answered correctly, and response times, measured as the time in seconds spent on each query. In Table 7 the mean accuracy, standard deviation, percentage of queries answered correctly, the standard error and a 95% confidence interval of both the fRAT items and the compound items are shown. The data for the compound items is additionally split up in comRAT-G items and Bowden & Jung-Beeman items. Table 8 shows the data on response times for queries that were answered correctly. In Table 9 the same measures are shown in regard to how many participants solved each query. The percentages here refer to the number of participants. Finally, Table 10 shows the descriptive statistics on how much time was spent on each query, independently of whether the correct solution was given or not.

16

Table 7: Descriptive statistics on accuracy, $n = 61$

	Mean no. of queries solved (SD), percentage	Std. Error	95% CI LB	95% CI UB
fRAT produced by comRAT-GF	17.1 (5.77), 35%	0.74	15.65	18.55
compound RAT	15.85 (7.6), 33%	0.97	13.95	17.75
– comRAT-G produced items	7.25 (3.72), 30%	0.48	6.31	8.18
– Bowden & Jung-Beeman items	8.61 (5.06), 25.9%	0.65	7.34	9.88

Table 8: Descriptive statistics on response times in seconds for correct answers, $n = 61$

	Mean response time in seconds (SD)	Std. Error	95% CI LB	95% CI UB
fRAT produced by comRAT-GF	14.14 (13.39)	1.71	10.78	17.5
compound RAT	11.68 (10.96)	1.42	8.89	14.48
– comRAT-G produced items	11.0 (10.62)	1.39	8.27	13.74
– Bowden & Jung-Beeman items	11.64 (10.65)	1.42	8.86	14.43

Table 9: Descriptive statistics on number of participants solving per query, $n = 61$

	Mean no. of participants solving (SD), percentage	Std. Error	95% CI LB	95% CI UB
fRAT produced by comRAT-GF	21.73 (16.65), 36%	2.13	12.47	20.83
compound RAT	20.15 (12.8), 33%	1.64	9.59	16.02
– comRAT-G produced items	18.42 (13.97), 30%	1.79	14.91	21.92
– Bowden & Jung-Beeman items	21.88 (11.56), 36%	1.48	18.97	24.78

Table 10: Descriptive statistics on mean time spent per query in seconds, $n = 61$

	Mean response time in seconds (SD)	Std. Error	95% CI LB	95% CI UB
fRAT produced by comRAT-GF	17.84 (6.6)	0.84	16.19	19.49
compound RAT	17.65 (4.57)	0.59	16.5	18.8
– comRAT-G produced items	18.57 (5.28)	0.68	17.25	19.9
– Bowden & Jung-Beeman items	16.83 (3.55)	0.45	15.94	17.72

In the verbal fluency tasks, participants named on average 43.75 words in the FAS paradigm ($SD = 13.89$, $SE = 1.78$, 95% CI [40.27, 47.24]) and 38.54 words in the categories paradigm ($SD = 12.55$, $SE = 1.61$, 95% CI [35.39, 41.7]).

375 As a measure of reliability, Cronbach's alpha was calculated for the fRAT-items, the compound items, and the comRAT-G items and Bowden & Jung-Beeman items individually, as well as for all items combined. This was done both for accuracy and response times, again independently of whether the right answer was given or not. The results can be found in Table 11.

Table 11: Cronbach's alpha for accuracy and response time

	Cronbach's alpha	95% CI LB	95% CI UB
Accuracy			
fRAT produced by comRAT-GF	0.79	0.72	0.86
compound RAT	0.87	0.82	0.91
– comRAT-G produced items	0.75	0.67	0.84
– Bowden & Jung-Beeman items	0.85	0.79	0.9
All items	0.88	0.83	0.92
RT			
fRAT produced by comRAT-GF	0.9	0.87	0.94
compound RAT	0.96	0.95	0.97
– comRAT-G produced items	0.93	0.9	0.95
– Bowden & Jung-Beeman items	0.92	0.89	0.95
All items	0.96	0.94	0.97

380 Accuracy of the functional and the compound items, as well as verbal fluency in the FAS and the category paradigm were all correlated with each other, as can be seen in Table 12. The response times of the functional and the compound items were correlated, showing a strong positive correlation which was highly significant ($r = .88, p < .001$).

385 Additionally, a strong and significant correlation was also found between the number of participants solving each functional query and the items' probability ($r = .71, p < .001$). The response time for each correctly solved query and its probability showed a negative correlation ($r = -.32, p < .05$), meaning that the higher the probability of an item, the less time is needed to come up with the correct solution. This means that a comRAT-F system solving the functional

RAT in the same way as comRAT-C solved the compound RAT would correlate in this way with human performace, on both Accuracy and Response Times.

Table 12: Correlations of accuracy of functional and compound items and verbal fluency

Variable	fRAT	Compound	FAS	Categories
fRAT	—	$r = .44, p < .001$	$r = .35, p < .05$	$r = .46, p < .001$
Compound		—	$r = .6, p < .001$	$r = .57, p < .001$
FAS			—	$r = .77, p < .001$
Categories				—

6. Discussion

In this section, we will discuss: (a) the suitability of the queries created with $comRAT-G_F$ via our current approach and (b) the results obtained in relation to verbal fluency and item probability.

6.1. Suitability of the approach for creating functional items

The results indicate that $comRAT-G_F$ can create functional RAT items. This shows that word associates are a decent seed foundation for the queries. As the performance of the human participants in functional items strongly correlates to that in the compound items (Accuracy $r = .44, p < 0.001$; RT $r = .88, p < 0.001$), and the items show a good level of internal validity (Cronbach alpha 0.79), we consider our approach of computationally generating functional RAT items from word associates successful.

Some of the queries generated from word associates also fulfill the criteria for compound items. The problem of this overlap can be resolved computationally by first generating compound items and then extracting them from the set of functional items. This would yield a dataset of functional associates without linguistic relations and exceed Worthen and Clark's constraints.

When analyzing the quality of the queries, the main constraints most probably stem from the dataset that was used as the basis and its suitability for the given problem: though a relatively good seed for our computational query

19

creation process, the South Florida association norms dataset [34] contains only one target word per cue word, and thus does not provide more distant associations which would be elicited later in the association flow. The queries generated based on this may be easy, since the first associate is most probably very close to the cue. So far, different difficulty levels were implemented using the frequency, defined as the number of people producing the target. Yet working with earlier and later associates as a difficulty modulator might add complexity and make the queries more interesting, since the different items might become conceptually more remote.

Another aspect that can be adressed are synonymy based relations between answer and query word. As these are easy to process, they might cause whole queries to be easier to solve, as for example Q1 in Table 2. What makes this query especially easy is the fact that all three query words are in the same semantic domain. Q9 on the other hand seems cognitively more demanding and also more interesing, since all cue words refer to different meanings of the answer word. Thus, when improving the quality of the queries based on synonymy, it will be beneficial to take the diversity of semantic domains in the cue words into account, instead of removing all synonymy based relations.

The relatedness between query words can also have an impact on the quality of the queries. Related queries do not require to merge different concepts and are thus easier. This can be tackled by checking all pairs (w_1, w_2), (w_2, w_3),(w_3, w_1) again with the knowledge base derived from the word associates norms and setting constraints for their relatedness. This would most probably not allow for queries such as Q2 to be generated, because $photo$ and $frame$ are closely related. Removing such adjacent relations will make the queries more complex and increase difficulty.

The approach explored here provides the advantage that functional RAT items can now be explored side by side with compound items. Knowing the frequency and the probability of such items can also help check various hypotheses on how the mechanism of remote association works, and the influence of these factors on it.

20

6.2. Relation to verbal fluency and probability of queries

Significant correlations were obtained between the two verbal fluency tasks used – the FAS and categorical fluency – and the performance in both compound and functional queries. The FAS correlates less with the functional ($r = .35$) than the compound ($r = .6$) RAT. The categorical fluency measure also correlates stronger with the functional ($r = .46$) than the compound ($r = .77$) RAT. Our exploratory hypothesis was that solving the FAS will somewhat be more related to compound skill, because it involved a search through known words, without a semantical component, while categorical fluency performance would correlate more with the functional RAT skill, because it involved a semantic search component. No indication that this might be so was obtained in this study. An experiment testing this hypothesis will be run in the future, setting up a semantically skilled versus a syntactically skilled group.

The first computational solver (comRAT-C) built to solve compound queries was compared to human performance in Bowden & Jung-Beeman's dataset. It was observed that the computational solver probability correlated at $r = .49$; $p < 0.002$ with participant accuracy and $r = -.52$; $p < 0.001$ with participant response times [26] – that is the higher the probability to solve, the less time participants took. This solver's mechanism was later used to computationally build the compound RAT [27] and now the functional query generator. The high significant correlation between item probability and human performance means that a comRAT-F system solving these queries based on the previously posited computational mechanism will correlate in its solving to human performance. This validates our previously posited computational mechanism, and our cognitive framework [29]. It indicates that these cognitive systems could be used in the future when aiming to predict human performance in RAT queries. It also shows that association strength, like word frequency (for compound queries), would have an impact in the ability to solve functional queries.

The correlation between the computationally generated compound items and Bowden & Jung-Beeman's set of items further validates the approach used to generate items in [27], with a different set of participants.

21

Finally, an interesting question refers to the limitations of extracting functional relationships from human data. On one hand, human produced associates may be of special interest in the building of cognitive systems because they provide a cognitive trace: for example, an earlier average production of some functional items as associates versus others may imply stronger connectivity to the initial given word. On the other hand, it is hard to ensure all functional relationships are extracted in such a manner. A possible approach, also suggested by one of our reviewers, would be to employ ontological hierarchies from WordNet [39, 40] to extract functional relations. However, such a source will not provide strength of relation between the word pairs. We plan to analyse the benefits of these approaches comparatively in the future.

As future work, we intend to: (a) improve queries by making sure no relations between the query words exist; (b) evaluate the interestingness of computationally created queries with human participants, as to improve the process of computational query creation and (c) computationally compare the benefits of different sources of functional associates in computational query creation.

Acknowledgements

[1] M. Boden, The Creative Mind: Myths and Mechanisms, Routledge, 2003.

[2] W. H. Batchelder, G. E. Alexander, Insight problem solving: A critical examination of the possibility of formal theory, The Journal of Problem Solving 5 (1) (2012) 56–100.

[3] K. Duncker, On problem solving, Psychological Monographs 58 (5, Whole No.270).

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

[5] K. H. Kim, Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT), Creativity research journal 18 (1) (2006) 3–14.

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

[7] R. Hass, M. Rivera, P. Silvia, On the dependability and feasibility of layperson ratings of divergent thinking 9.

[8] E. Threadgold, J. Marsh, L. Ball, Normative data for 84 uk english rebus puzzles, Frontiers in Psychology 9 (2018) 1–15. `doi:10.3389/fpsyg.2018.02513`.

[9] R. Hass, R. Beaty, Use or consequences: Probing the cognitive difference between two measures of divergent thinking, Frontiers in Psychology 9 (2018) 2327. `doi:10.3389/fpsyg.2018.02327`.

[10] S. A. Mednick, M. Mednick, Remote associates test: Examiner's manual, Houghton Mifflin, 1971.

[11] A.-M. Olteţeanu, B. Gautam, Z. Falomir, 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, CEUR-Ws, 2015, pp. 19–28.

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

[13] J. B. Cunningham, J. N. MacGregor, J. Gibb, J. Haar, Categories of insight and their correlates: An exploration of relationships among classic-type insight problems, rebus puzzles, remote associates and esoteric analogies, The Journal of Creative Behavior 43 (4) (2009) 262–280.

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

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

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

[17] C. Salvi, G. Costantini, E. Bricolo, M. Perugini, M. Beeman, Validation of italian rebus puzzles and compound remote associate problems, Behavior research methods 48 (2) (2016) 664–685.

[18] S. A. Chermahini, M. Hickendorff, B. Hommel, Development and validity of a Dutch version of the Remote Associates Task: An item-response theory approach, Thinking Skills and Creativity 7 (3) (2012) 177–186.

[19] N. Landmann, M. Kuhn, H. Piosczyk, B. Feige, D. Riemann, C. Nissen, Entwicklung von 130 deutschsprachigen compound remote associate (cra)-worträtseln zur untersuchung kreativer prozesse im deutschen sprachraum, Psychologische Rundschau.

[20] Y. Baba, An analysis of creativity by means of the Remote Associates Test for Adult Revised in Japanese (JARAT FORM A)., Japanese Journal of Psychology.

[21] W. Shen, Y. Yuan, C. Liu, B. Yi, K. Dou, The development and validity of a chinese version of the compound remote associates test, American Journal of Psychology 129 (3) (2016) 245–258.

[22] A. Sobków, A. Połeć, C. Nosal, RAT-PL–construction and validation of Polish version of remote associates test, Studia Psychologiczne 54 (2) (2016) 1–13.

[23] S. Mednick, The associative basis of the creative process., Psychological review 69 (3) (1962) 220.

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

[25] D. S. Palermo, J. J. Jenkins, Word association norms: Grade school through college.

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

[27] A.-M. Olteţeanu, H. Schultheis, J. B. Dyer, Computationally constructing a repository of compound Remote Associates Test items in American English with comRAT-G, Behavior Research Methods, Instruments, & Computers `doi:10.3758/s13428-017-0965-8`.

[28] A.-M. Olteţeanu, From simple machines to Eureka in four not-so-easy steps.Towards creative visuospatial intelligence, in: V. Müller (Ed.), Fundamental Issues of Artificial Intelligence, Vol. 376 of Synthese Library, Springer, 2016, pp. 159–180. `doi:10.1007/978-3-319-26485-1_11`.

[29] A.-M. Olteţeanu, Two general classes in creative problem-solving? An account based on the cognitive processes involved in the problem structure - representation structure relationship., in: T. Besold, K.-U. Kühnberger, M. Schorlemmer, A. Smaill (Eds.), Proceedings of the International Conference on Computational Creativity, Vol. 01-2014 of Publications of the Institute of Cognitive Science, Osnabrück, 2014.

25

[30] A. Augello, I. Infantino, A. Lieto, G. Pilato, R. Rizzo, F. Vella, Artwork creation by a cognitive architecture integrating computational creativity and dual process approaches, Biologically Inspired Cognitive Architectures 15. `doi:10.1016/j.bica.2015.09.007`.

[31] G. A. Wiggins, Creativity, information, and consciousness: The information dynamics of thinking, Physics of Life Reviews`doi:https://doi.org/10.1016/j.plrev.2018.05.001`.
URL `http://www.sciencedirect.com/science/article/pii/S1571064518300599`

[32] A.-M. Olteţeanu, Z. Falomir, C. Freksa, Artificial cognitive systems that can answer human creativity tests: An approach and two case studies, IEEE Transactions on Cognitive and Developmental Systems 10 (2) (2018) 469–475.

[33] V. Mekern, B. Hommel, Z. Sjoerds, Computational models of creativity: a review of single-process and multi-process recent approaches to demystify creative cognition, Current Opinion in Behavioral Sciences 27 (2019) 47 – 54, creativity. `doi:https://doi.org/10.1016/j.cobeha.2018.09.008`.
URL `http://www.sciencedirect.com/science/article/pii/S2352154618301256`

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

[35] A.-M. Olteteanu, K. Yoopoo, Towards computationally creating multi-answer queries for the remote associates test, in: Proceedings of the 5th International Workshop on Artificial Intelligence and Cognition, Vol. 2090, CEUR-Ws, 2017, pp. 34–40.
URL `http://ceur-ws.org/Vol-2090/`

[36] A.-M. Olteţeanu, H. Schultheis, What determines creative association? revealing two factors which separately influence the creative process when

solving the remote associates test, The Journal of Creative Behaviour`doi:`
`10.1002/jocb.177`.

[37] A. Tucker, Applied combinatorics, John Wiley&Sons, Inc., 2006.

615 [38] G. V. Georgiev, D. D. Georgiev, Enhancing user creativity: Semantic measures for idea generation, Knowledge-Based Systems 151 (2018) 1–15.

[39] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, Introduction to WordNet: An on-line lexical database, International journal of lexicography 3 (4) (1990) 235–244.

620 [40] C. Fellbaum, 1998, WordNet: An electronic lexical database.

Appendix

Table 13: Accuracy and reaction times (in seconds) for correctly solved items. Items evaluated with human participants in study II. w_{ans} stands for the answer word, while $w_1 - w_3$ stand for the query items.

Query no.	w_1	w_2	w_3	w_{ans}	mean accuracy	SD accuracy	mean RT	SD RT
1	question	reply	solution	answer	0.52	0.5	9.1	6.1
2	sensitive	sob	weep	cry	0.51	0.5	11.9	13.7
3	antlers	doe	fawn	deer	0.89	0.32	13.1	29.4
4	bud	dandelion	petals	flower	0.89	0.32	9.3	7.4
5	colt	mare	unicorn	horse	0.7	0.46	8.6	13.9
6	crown	royaly	throne	king	0.48	0.	16.8	15.6
7	algebra	calculus	trigonometry	math	0.92	0.28	6.9	4.5
8	pedal	pull	shove	push	0.33	0.47	11.7	7.1
9	clockwise	left	wrong	right	0.13	0.34	9.5	7.8
10	flu	nauseous	virus	sick	0.57	0.5	19.6	72.9
11	astronomy	moon	twinkle	star	0.67	0.47	10.1	10.7
12	bait	pond	tuna	fish	0.93	0.25	9.7	11.1
13	bandaid	trim	wound	cut	0.16	0.37	20.0	19.0
14	gravity	low	uo	down	0.11	0.32	10.9	4.7
15	emergency	rapid	slow	fast	0.08	0.28	10.7	8.2
16	brawl	debate	soldier	fight	0.51	0.5	11.2	9.5
17	birds	frog	kite	fly	0.46	0.5	14.9	12.1
18	finger	glove	palm	hand	0.72	0.45	9.8	9.9
19	bed	darkness	sedative	sleep	0.59	0.5	10.1	12.4
20	discuss	gossip	telephone	talk	0.54	0.5	10.9	13.1
21	fangs	gums	wolf	teeth	0.62	0.49	11.1	10.0
22	marsh	saliva	slippery	wet	0.48	0.5	10.8	10.3
23	dictionary	verse	vocabulary	words	0.44	0.5	12.1	19.6
24	fault	incorrect	unjust	wrong	0.41	0.5	9.2	6.1
25	murder	operate	vein	blood	0.3	0.46	15.0	17.5
26	empire	moat	princess	castle	0.59	0.5	13.2	12.2
27	bench	sofa	stool	chair	0.08	0.28	21.3	19.8
28	beaker	flask	science	chemistry	0.13	0.34	9.5	6.8
29	adults	development	yo-yo	children	0.03	0.18	8.2	0.8
30	cemetery	coma	noose	dead	0.18	0.39	62.8	171.2
31	exam	scare	terror	fear	0.15	0.36	18.2	8.5
32	hand	toe	trigger	finger	0.36	0.48	21.3	21.0
33	angel	church	faith	god	0.13	0.34	9.0	6.1
34	body	commander	scull	head	0.28	0.45	12.4	8.2
35	cello	scalpel	trumpet	instrument	0.61	0.49	28.3	74.0
36	desk	quill	stapler	pen	0.18	39	16.6	14.4
37	arrest	badge	deputy	cop	0.08	0.28	4.1	0.4
38	electron	inertia	zest	energy	0.13	0.34	10.8	7.9
39	diet	strain	sweat	exercise	0.31	0.47	20.6	18.1
40	assault	cop	murder	gun	0.07	0.25	13.7	10.1
41	drill	grave	spike	hole	0.05	0.22	11.3	5.0
42	care	tactful	willing	kind	0.0	0.0	–	–
43	midnight	saturn	wolf	moon	0.43	0.5	14.4	19.5
44	bloom	opportunity	split	open	0.05	0.22	17.2	2.2
45	accomplished	dolphin	sly	smart	0.15	0.36	24.6	20.7
46	duck	sardine	sinker	swim	0.02	0.13	22.3	0.0
47	europe	mushroom	pack	trip	0.0	0.0	–	–
48	fierce	steel	warrior	strong	0.13	0.34	11.1	10.9