Are all Remote Associates Test equal?
An overview and comparison of the Remote Associates Test in different languages

Jan Philipp Behrens (jan.behrens@fu-berlin.de)
Cognitive Systems Group, Human-Centered Computation
Freie Universität Berlin, Germany

Ana-Maria Olteţeanu (ana-maria.olteteanu@fu-berlin.de)
Cognitive Systems Group, Human-Centered Computation
Freie Universität Berlin, Germany

Abstract

The Remote Associates Test (RAT, CRA) is a classical creativity test used to measure creativity as a function of associative ability. The RAT has been administered in different languages. Nonetheless, because of how embedded in the language the test is, only a few items are directly translatable, and most of the time the RAT is created anew in each language. This process of manual (and in two cases computational) creation of RAT items is guided by the researchers' understanding of the task. However, are the RAT items in different languages comparable? In this paper, different RAT stimuli datasets are analyzed qualitatively and quantitatively. Significant differences are observed between certain datasets in terms of solver performance. The potential sources of these differences are discussed, together with what this means for creativity psychometrics and computational vs. manual creation of stimuli.

Keywords: Remote Associates Test; RAT; CRA; Creativity; Creativity evaluation and metrics; Creativity Test

Introduction

The Remote Associates Test is a creativity test often used in the literature (Ansburg & Hill, 2003; Cunningham, MacGregor, Gibb, & Haar, 2009; Mednick & Mednick, 1971; Cai, Mednick, Harrison, Kanady, & Mednick, 2009; Ward, Thompson-Lake, Ely, & Kaminski, 2008).
A RAT problem given to a participant contains three words, for example FISH, MINE, RUSH; the participant has to come up with a fourth word related to all of the three given words. In this case, GOLD is an answer, because the compounds GOLD FISH, GOLD MINE, GOLD RUSH can be built with it. For a human or a machine (Olteţeanu & Falomir, 2015) to solve the RAT, knowledge about the compound words of a language is needed.

Because solving the RAT relies on knowing various expressions and compound words from a language, native speakers have an advantage and are generally the target population when deploying the RAT. This raises the need for various RAT stimuli sets in different languages.

As the RAT relies on knowledge and expressions which are language specific, the RAT is, in most part, not translatable between languages. An exception to this are the rare cases in which all compounds required as knowledge by a RAT item in a specific language also exist in another language - for example GOLDFISCH, GOLDMINE, GOLDRAUSCH as the German counterpart of the above mentioned query.

As only a few items are translatable, RAT sets of items are created anew by researchers in each language. This entails that RAT queries are probably impacted by the language itself, and quite likely by the preferences and knowledge of compound words of the stimuli dataset authors. The Remote Associates Test (RAT) is administered in many creativity studies, in the native language of the participants. Results reported in these studies are therefore impacted by the quality and difficulty of RAT items in each language. How can this impact be assessed?

No overview exists of the human performance in the RAT / CRA in the different languages. Such an overview would help us understand whether significant differences exist between performance on different RAT problem sets in the various languages in which it is employed. If no significant differences exist, this may indicate that results reported on creativity studies which use the RAT in different languages are, indeed, cross-comparable. If a significant difference however does exist, the comparability of the RAT across languages may require more nuance, and the development of an understanding of the sources of this difference.

This paper sets out to construct an overview of the RAT across eight languages and two types of the RAT (compound and functional), and provide an initial analysis between RAT sets across all these languages.

The RAT and languages

Sets of RAT / CRA problems of the following languages were analyzed - please note that some languages come with multiple datasets (D):

- German (Landmann et al., 2014)
- Chinese (Shen, Yuan, Liu, Yi, & Dou, 2016)
- Italian (Salvi, Costantini, Bricolo, Perugini, & Beeman, 2016)
- Romanian (Olteţeanu, Taranu, & Ionescu, n.d.)
- Polish (Sobków, Połeć, & Nosal, 2016)
- English D1 (Bowden & Jung-Beeman, 2003)
- English D2 (Olteţeanu, Schultheis, & Dyer, 2017)
- English D3 (Olteţeanu, Schöttner, & Schuberth, 2019)

– Finnish (Toivainen, Olteţeanu, Repeykova, Likhanov, & Kovas, 2019)

– Russian (Toivainen et al., 2019)

RAT comparison

A qualitative and quantitative comparison of the RAT datasets above is provided in the next sections.

Qualitative comparison

English datasets D2 and D3 contain different types of items: *compound* versus *functional*. For compound items, the relationship between the three given words and the answer word is a relationship manifested in language – for example, GOLD FISH, GOLD MINE and GOLD RUSH are compounds which all appear in language. By contrast, the relationship between functional query words and the answer reflects a functional relationship between the two, but may or may not be a compound linguistic relationship. For example, the relationship between CLOCKWISE and RIGHT or WRONG and RIGHT is a functional relationship. Of the above datasets, English D3 is functional.

Independent of the compound/functional classification, RAT problems have also been divided into two types based on the order of the words: homogeneous and heterogeneous items. RAT items are homogeneous if the solution word is either a prefix or a suffix to all the three words of the problem (like in the query FISH, MINE, RUSH, where GOLD acts as a prefix to each of the query items). Problems are heterogeneous, if the solution word is the prefix for some of the words and a suffix to other words of the problem (e.g. in the query RIVER, NOTE, ACCOUNT, the answer BANK is a suffix for the first word, and a prefix for the other two).

Of the above datasets, the German, Italian and English D1 ones distinguished between the heterogenous and homogeneous type of the queries. ANOVA with task type as a factor were run by the authors on these sets. The task type factor showed no significant effect on Accuracy (the number of queries solved by the participants). In the German version, a significant effect of the task type factor was observed on reaction times.

Finally, of the dataset items above, most are manually created. An exception to this are items from the English D2 and English D3 datasets. English D2 (Olteţeanu et al., 2017) successfully attempts the computational creation of RAT items, and compares results with an existing (English D1) normative dataset. English D3 (Olteţeanu et al., 2019) applies the computational approach using a new type of language knowledge to the creation of functional items, thus resurrecting an older idea of Worthen and Clark (Worthen & Clark, 1971) regarding the existence of such items, and their differences to compound items. These items are compared to compound items in the paper.

Quantitative comparison

In the following, a descriptive statistics overview of the different datasets is provided. To answer the question whether differences exist between RAT datasets in the various languages, Welch's unequal variances t-test is used on each two language pairs to determine the effect of language on the Remote Associates Test.

Descriptive data

The various RAT datasets contained varying numbers of items, between 17 (Polish) and 144 (English D1). Furthermore, the various items were deployed either (a) giving participants different timeframes to solve each query, between 2s and 60s, or (b) without setting a time limit. Since 2s, 5s and 7s timeframes were only used once across these datasets, only items between 15s and 60s are analysed in this paper. The stimuli were deployed on populations of various sizes, with n ranging between 26 and 317 participants. The Accuracy (number of correct answers given by the participants) fluctuated between .31 and .58. The response times ranged between $7.26s$ and $37.34s$. Please note that means and standard deviations were calculated for this paper from the given data, where they were not provided by the initial dataset. Table 4 gives an overview of all the datasets and various descriptive metrics across all languages.

Cronbach's alpha

Cronbach's alpha is the most commonly used method for estimating the reliability of a test, as reflected by its internal consistency between items. Scores below 0.5 indicate an unacceptable internal consistency, whereas higher scores indicate a better one. Generally scores above 0.7 are considered to reflect an acceptable amount of reliability, and an α above 0.9 is excellent. The Cronbach α scores were calculated by authors for some of the intial papers (see Table 4) and vary between .73 and .99.

Differences between languages

In order to measure differences between languages, heterogenous and homogeneous items were combined and Welch's unequal variances t-test was conducted to measure the difference between means on two existing performance metrics: Accuracy and Response Times.

Accuracy in 15s timeframe

As shown in Table 1, there were significant differences of means between:

- Italian ($M = .39$; $SD = .23$) and German ($M = .30$; $SD = .27$); $t(250) = 2.86$, $p = .0046$
- Italian and English D1 ($M = .31$; $SD = .22$); $t(253.88) = 2.95$, $p = .0035$

Table 1: Welch test results for accuracy in a 15s timeframe

accuracy	GER			ITA		
15s	t	df	p	t	df	p
ITA	2.86	249.99	.005**	-	-	-
ENG D1	0.13	260.92	.89	2.95	253.88	.004**

Accuracy in 30s timeframe

Like displayed in Table 5, there were significant differences of means between:

- Chinese ($M = .58$; $SD = .25$) and Polish ($M = .41$; $SD = .23$); $t(38.29) = 4.92$, $p < .0001$
- Chinese and German ($M = .30$; $SD = .27$); $t(254.28) = 5.92$, $p < .0001$
- Chinese and English D1 ($M = .31$; $SD = .22$); $t(265.86) = 3.47$, $p = .0006$
- English D1 and German; $t(262.27) = 2.72$, $p = .007$

Accuracy without timeframe

As reported in Table 6, there were significant differences of means between:

- English D2 ($M = .52$; $SD = .14$) and Finnish ($M = .46$; $SD = .11$); $t(93.95) = 2.1$, $p.038$
- English D3 ($M = .33$; $SD = .16$) and Romanian ($M = .54$; $SD = .43$); $t(83.26) = 3.98$, $p = .0002$
- English D3 and Russian ($M = .55$; $SD = .14$); $t(92.87) = 3.73$, $p = .0003$
- English D3 and English D2; $t(93.46) = 3.83$, $p = .0002$

RT in 15s timeframe

As presented in Table 2, there was a significant difference of means between:

- English D1 ($M = 7.26$; $SD = 1.65$) and Italian ($M = 6.52$; $SD = 1.46$); $t(258.86) = 3.87$, $p = .0001$.

RT in 30s timeframe

As shown in Table 3, there were significant differences of means between:

- English D1 ($M = 10.45$; $SD = 3.47$) and Polish ($M = 14.03$; $SD = 3.06$); $t(21.38) = 4.48$, $p = .0002$
- Chinese ($M = 9.74$; $SD = 3.13$) and Polish; $t(20.7) = 5.42$, $p < .0001$

RT without timeframe

As stated in Table 7, there were significant differences of means between:

- Finnish ($M = 37.34$; $SD = 17.36$) and Romanian ($M = 15.37$; $SD = 10.53$); $t(52.72) = 6.67$, $p < .0001$
- Finnish and Russian ($M = 23.53$; $SD = 10.38$); $t(58.18) = 5.05$, $p < .0001$
- Finnish and English D2 ($M = 14.52$; $SD = 9.89$); $t(76.07) = 4.79$, $p < .0001$
- Finnish and English D3 ($M = 11.68$; $SD = 10.96$); $t(67.26) = 6.48$, $p < .0001$
- Russian and English D3; $t(83.71) = 2.99$, $p = .004$
- Russian and Romanian; $t(91.38) = 3.37$, $p = .001$
- English D2 and English D3; $t(91.92) = 2.09$, $p = .04$

Discussion and further work

The hardest sets to solve seem to be the English D3 set of items from Study 2, with a an average accuracy of .30, and the Finnish dataset in terms of response times, with a mean 37.34 seconds. The response times of the Russian RAT were also noticeably higher (23.53s).

This paper set out to compare the RAT in different languages, and across different datasets. Significant differences were observed between multiple languages and datasets, on both the Accuracy and Response Times performance metrics.

The significant difference observed between the English D2 and English D3 sets may have as a source the difference between types of items (compound versus functional).

In the cases in which a significant difference exists between different language datasets, various causes could act as the source:

(a) different population samples are more creative (or at least better at the associative factor in creativity);

(b) the RAT is more difficult in some languages, because of the language itself and the cognitive factors resulting from encoding linguistic knowledge and solving the RAT in that language and/or

(c) sets of RAT queries vary in difficulty, because they are created without using standardized methods, thus depend on the inspiration and knowledge base of the researchers creating them.

This initial investigation shows that differences between the RAT in various languages need to be addressed in more detail. Before cross-comparison of creativity results can be declared, the source of these differences needs to be found. Experimental or analytical setups need to be designed in order to establish which one of (a), (b) and (c), or combination thereof, is the source of the differences.

An initial thought on establishing comparability could be to attempt to find translatable items across the various languages. By keeping stimuli items constant, differences of creativity pertaining to the population or use of language could be established.

However, even if translatable, the same RAT items may not be the same difficulty in different languages. Some light on this is shed by computational models like comRAT-C (Olteţeanu & Falomir, 2015), essentially models of memory search, which can solve the RAT by organizing their knowledge in a semantic net-like structure and propagating activation through word associations. The comRAT-C's probability

of solving a query correlates with human performance. This model entails that, even if different RAT queries can be translated in different languages, equivalence does not necessarily exist between them: the number of word associates and the strength of association may not be the same in different languages. Different tools may thus need to be used to try to establish query equivalence.

A potential solution may be to establish a stronger item equivalence in computational terms: for example by using computational RAT query generators like comRAT-G (Olteţeanu et al., 2017), to create sets of items where a high degree of control can be maintained over the number of associates and the association strength of the query words. Such approaches have already proven fruitful in the deployment of more precise empirical designs (Olteţeanu & Schultheis, 2017), and in the creation of other types of items (Olteţeanu et al., 2019).

Another direction of future work would be to establish a creative association measure which transcends the constraints of language like a visual Remote Associates Test - some work in this direction has already been done by (Olteţeanu, Gautam, & Falomir, 2015; Toivainen et al., 2019).

This paper gives an overview of RAT datasets in multiple languages, and shows that cross-linguistic comparability should not be taken for granted in the case of this broadly used creativity test.

Acknowledgements

The support of the Deutsche Forschungsgemeinschaft (DFG) for the project CreaCogs via grant OL 518/1-1 is gratefully acknowledged.

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.

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

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. , *106*(25), 10130–10134.

Cunningham, J. B., MacGregor, J., Gibb, J., & Haar, J. (2009). 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), 262-280.

Landmann, N., Kuhn, M., Piosczyk, H., Feige, B., Riemann, D., & Nissen, C. (2014). Entwicklung von 130 deutsch sprachigen Compound Remote Associate (CRA)-Wortraetseln zur Untersuchung kreativer Prozesse im deutschen Sprachraum. *Psychologische Rundschau*, *65*, 200–211.

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

Olteţeanu, A.-M., & Falomir, Z. (2015). comrat-c : A computational compound remote associate test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, *67*, 81-90.

Olteţeanu, A.-M., Gautam, B., & Falomir, Z. (2015). Towards a visual remote associates test and its computational solver..

Olteţeanu, A.-M., Schöttner, M., & Schuberth, S. (2019). Computationally resurrecting the functional remote associates test using cognitive word associates and principles from a computational solver. *Knowledge-Based Systems*.

Olteţeanu, A.-M., & Schultheis, H. (2017). What determines creative association? Revealing two factors which separately influence the creative process when solving the remote associates test. *The Journal of creative behavior*. doi: 10.1002/jocb.177

Olteţeanu, A.-M., Schultheis, H., & Dyer, J. B. (2017). Computationally constructing a repository of compound Remote Associates Test items in American English with comRAT-G. *Behavior Research Methods*, *7*.

Olteţeanu, A.-M., Taranu, M., & Ionescu, T. (n.d.). Normative data for 111 compound Remote Associates Test problems in Romanian. *Frontiers*.

Salvi, C., Costantini, G., Bricolo, E., Perugini, M., & Beeman, M. (2016). Validation of Italian rebus puzzles and compound remote associate problems. *Behavior Research Methods*, *48*, 664–685.

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

Sobków, A., Połeć, A., & Nosal, C. (2016). Rat-pl – construction and validation of polish version of remote associates test. *Studia Psychologiczne*, *54*, 1–13.

Toivainen, T., Olteţeanu, A.-M., Repeykova, V., Likhanov, M., & Kovas, Y. (2019). Visual and linguistic stimuli in the remote associates test: A cross-cultural investigation. *Frontiers in Psychology*, *10*.

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.

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

Appendix

Table 2: Welch test results for RT in a 15s timeframe

RT	ITA		
15s	t	df	p
ENG D1	3.87	258.86	.0001***

Table 3: Welch test results for RT in a 30s timeframe

RT	ENG D1			CHI D1		
30s	t	df	p	t	df	p
CHI D1	1.77	265.91	.08	-	-	-
POL	4.48	21.38	.0002***	5.42	20.7	2e-5****

Table 4: Number of elements($|x|$), sample size(n), mean($\bar{x}$) and standard deviation(s) of accuracy and response time and Cronbach's α for the RAT in the different languages. S1 and S2 reflect different studies using the same set of stimuli.

	Timeframe			Accuracy $\bar{x}$ (s)		RT $\bar{x}$ (s)	Cronbach's α	
Language	in sec	$\|x\|$	n	sum	per item	per item [sec]	Accu	RT
German both	60	130	80	54.99 (34.97)	.44 (.27)	16.97 (7.12)	-	-
heterogeneous	60	56	80	26.10 (15.79)	.47 (.28)	15.80 (6.70)	-	-
homogeneous	60	74	80	30.19 (19.17)	.41 (.26)	18.50 (7.50)	-	-
German both	30	130	80	-	.39 (.27)	-	-	-
German both	15	130	80	-	.30 (.27)	-	-	-
Chinese	30	128	123	74.46	.58 (.25)	9.74 (3.13)	.92	
Italian both	15	122	317	47.58 (28.06)	.39 (.23)	6.52 (1.46)	-	-
heterogeneous	15	66	317	25.48 (14.72)	.39 (.22)	-	-	-
homogeneous	15	56	317	22.12 (13.44)	.40 (.24)	-	-	-
Romanian	none	111	63	59.94 (47.73)	.54 (.43)	15.37 (10.53)	.93	.97
Polish	30	17	206	6.90 (3.90)	.41 (.23)	14.02 (3.06)	.79	-
English D1 both	30	144	289	72.72	.51 (.25)	10.45 (3.47)	-	-
heterogeneous	30	59	289	29.74	.50	-	-	-
homogeneous	30	85	289	42.93	.51	-	-	-
English D1 both	15	144	289	-	.31 (.22)	7.26 (1.65)	-	-
English D2 both	none	100	113	52.64 (16.16)	.53 (.16)	-	.94	.99
comRAT-G	none	50	113	26.20 (7.03)	.52 (.14)	14.52 (9.89)	.85	.99
Bowden, J.-B.	none	50	113	26.41 (11.24)	.53 (.23)	16.56 (12.84)	.93	.99
English D3 S1 fRAT	none	75	26	35.27 (7.99)	.47 (.11)	13.91 (8.42)	-	-
comRAT	none	50	26	25.02 (7.26)	.50 (.15)	12.38 (6.23)	-	-
English D3 S2 fRAT	none	48	61	17.10 (5.77)	.36 (.12)	14.14 (13.39)	.79	.90
Compound both	none	48	61	15.85 (7.60)	.33 (.16)	11.68 (10.96)	.87	.96
comRAT-G	none	24	61	7.25 (3.72)	.30 (.16)	11.00 (10.62)	.75	.93
Bowden, J.-B.	none	24	61	8.61 (5.06)	.36 (.21)	11.64 (0.65)	.85	.92
Finnish	none	47	67	21.60 (5.30)	.46 (.11)	37.34 (17.36)	.73	-
Russian	none	48	67	26.60 (6.90)	.55 (.14)	23.53 (10.38)	.83	-

Table 5: Welch test results for accuracy in a 30s timeframe

accuracy	GER			CHI D1			POL		
30s	t	df	p	t	df	p	t	df	p
CHI D1	5.92	254.28	1e-8****	-	-	-	-	-	-
POL	0.39	43.32	.7	4.92	38.29	2e-5****	-	-	-
ENG D1	2.72	262.27	.007**	3.47	265.86	.0006***	2.03	36.62	.05

Table 6: Welch test results for accuracy without timeframe

accuracy	ROM			FIN			RUS			ENG D2		
no tf	t	df	p	t	df	p	t	df	p	t	df	p
FIN	1.66	78.25	.10	-	-	-	-	-	-	-	-	-
RUS	0.32	91.93	.75	1.71	90.40	.09	-	-	-	-	-	-
ENG D2	1.00	74.86	.32	2.10	93.95	.038*	0.66	89.52	.51	-	-	-
ENG D3	3.98	83.26	.0002***	1.82	92.66	.072	3.73	92.87	.0003***	3.83	93.46	.0002***

Table 7: Welch test results for response time without a timeframe

RT	ROM			FIN			RUS			ENG D2		
no tf	t	df	p	t	df	p	t	df	p	t	df	p
FIN	6.67	52.72	2e-8****	-	-	-	-	-	-	-	-	-
RUS	3.37	91.38	.001**	5.05	58.18	5e-6****	-	-	-	-	-	-
ENG D2	1.96	66.05	.054	4.79	76.07	8e-6****	0.30	80.50	.76	-	-	-
ENG D3	0.64	68.42	.52	6.48	67.26	1e-8****	2.99	83.71	.004**	2.09	91.92	.04*