frontiers in Psychology

ORIGINAL RESEARCH
published: 30 June 2020
doi: 10.3389/fpsyg.2020.01125

Are All Remote Associates Tests Equal? An Overview of the Remote Associates Test in Different Languages

Jan Philipp Behrens and Ana-Maria Olteţeanu**

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

The Remote Associates Test (RAT, CRA) is a classic creativity test used to measure creativity as a function of associative ability. The RAT has been administered in various different languages. Nonetheless, because of how embedded in language the test is, only a few items are directly translatable, and most of the time, the RAT is created a new 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. This paper focuses on the question of whether RAT datasets administered in different languages within the literature are comparable. To answer this question, datasets acquired using different RAT stimuli are analyzed qualitatively and quantitatively. Kruskal-Wallis tests are conducted to find out whether there is a significant difference between any of the datasets for a given time frame. Pairwise Mann-Whitney *post-hoc* tests are then used to find out which pairs are different. Significant differences are observed between 18 dataset pairings regarding Accuracy and between 16 in terms of Response Time. 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

OPEN ACCESS

Edited by:
Eddy J. Davelaar, Birkbeck, University of London, United Kingdom

Reviewed by:
Edward Bowden, University of Wisconsin-Parkside, United States
Caroline Di Bernardi Luft, Queen Mary University of London, United Kingdom

****Correspondence:***
Jan Philipp Behrens
jabehrens@uni-potsdam.de
Ana-Maria Olteţeanu
ana-maria.olteteanu@fu-berlin.de

Specialty section:
This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology

Received: *02 June 2019*
Accepted: *04 May 2020*
Published: *30 June 2020*

Citation:
Behrens JP and Olteţeanu A-M (2020) Are All Remote Associates Tests Equal? An Overview of the Remote Associates Test in Different Languages. Front. Psychol. 11:1125. doi: 10.3389/fpsyg.2020.01125

1. INTRODUCTION

The Remote Associates Test is a creativity test that is often used in the literature (Mednick and Mednick, 1971; Ansburg and Hill, 2003; Ward et al., 2008; Cai et al., 2009; Cunningham et al., 2009). 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 GOLDFISH, GOLD MINE, GOLD RUSH can be built with it. For a human or a machine (Olteţeanu and 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 gives rise to a need for different RAT stimulus sets in different languages.

As the RAT relies on knowledge and expressions that are language-specific, the RAT is, in most part, not translatable between languages. Exceptions 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 in each language by researchers. This means that RAT queries are probably impacted by the language itself and quite likely by the preferences and knowledge of compound words of the authors of the stimulus dataset. The Remote Associates Test (RAT) in the native language of the participants is administered in many creativity studies. 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 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 for creativity studies that use the RAT in different languages are, indeed, cross-comparable. If a significant difference does exist, however, the comparability of the RAT as a tool 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 RAT (compound and functional) and to provide an initial comparative analysis between RAT sets across all of these languages. Section 2 introduces the different language datasets that will be used. The third section compares the RAT datasets quantitatively and qualitatively. In section 4, results are presented regarding the differences between language and gender. The fifth and last section discusses the results and gives a view of possible future work.

2. THE REMOTE ASSOCIATES TEST AND LANGUAGES

Sets of RAT/CRA problems in the following languages were analyzed—please note that some languages have multiple datasets (D):

- German (Landmann et al., 2014)
- Chinese D1 (Shen et al., 2016)
- Chinese D2 (Wu and Chen, 2017)
- Italian (Salvi et al., 2016)
- Romanian (Olteţeanu et al., 2019b)
- Polish (Sobków et al., 2016)
- English D1 (Bowden and Jung-Beeman, 2003)
- English D2 (Olteţeanu et al., 2017)
- English D3 (Olteţeanu et al., 2019a)
- Finnish (Toivainen et al., 2019)
- Russian (Toivainen et al., 2019)

The Dutch (Chermahini et al., 2012) and both of the Japanese versions (Baba, 1982) and (Orita et al., 2018) had to be excluded because either the author was unreachable or the requested data were not sent to us in time.

3. REMOTE ASSOCIATES TEST COMPARISON

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

3.1. Qualitative Comparison

English datasets D2 and D3 contain different types of items: *compound* vs. *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 that all appear in language. By contrast, the relationship between functional query words and the answer reflects a functional relationship between these words, and there 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 three of the words in 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 the suffix to other words in 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 distinguish between heterogeneous and homogeneous queries. ANOVAs with task type as a factor were run by the dataset authors on these sets. The task-type factor showed no significant effect on accuracy (the number of queries solved by the participants). Only in the German version was a significant effect of the task-type factor on reaction times observed.

Because of the linguistic differences between Chinese and English, the Chinese authors came up with a character pairing method rather than compound words. In the authors' example, 生(to generate), 天(the sky), and 温(warm) paired with the solution creates three actual two-character words. The answer, in this case, would be 氣(air), and the resulting two-character words are 生氣(anger), 天氣(weather), and 氣温(temperature).

The Chinese D2 distinguished not between *heterogenous* and *homogeneous* but between *heteronym* and *non-heteronym* words. A heteronym is a word that has the same spelling but different pronunciation and meaning, for example, desert (arid region)/desert (leave). They found that the pass rate on heteronymous items was lower for the 20 and 30-s time limit condition but that the response time was not, indicating that heteronymous items were more difficult.

3.1.1. Test Item Creation

In the Italian study, 150 CRA items inspired by Mednick (1962) were initially tested and then reduced to 122 items by filtering out items that were always or never solved. At the beginning, the German study also contained 150 items. Its creation was based

on the original of Bowden and Jung-Beeman (2003) and was later filtered down to 130 items because 13 items had multiple solutions and 7 contained unclear words. The approach of the Romanian study was to first translate items from Bowden and Jung-Beeman (2003) and Salvi et al. (2016). If the translation was impossible (most items), the item was adapted or a single translated word out of the item was used as a seed for the creation of a new item. Afterward, the 198 created items were rated by the authors and five student volunteers in terms of how suitable they were, and then the dataset was reduced to the 111 most suitable items. The Polish dataset was created based on the original items of Bowden and Jung-Beeman (2003) and first contained 50 triads. These were then further reduced to 25 with diverse difficulty and one dominating solution. A subsequent test resulted in another reduction to 17 triads because of low factor loadings. The 47 Finnish items were all created by the research team, whereas the 48 Russian ones contained 12 created items and 36 items adopted from Druzhinin (1999). The authors of the Chinese D1 selected, according to Sio and Rudowicz (2007), 192 out of 288 items previously constructed by Jen et al. (2004), with the criterion that no solutions were repeated or used as problem words. After another reduction based on relative difficulty, the dataset consisted of 128 items. The Chinese D2 authors designed 120 items based on Mednick and Mednick (1971), Bowden and Jung-Beeman (2003), and Jen et al. (2004), of which they finally used the 90 that had a pass rate above 0%.

Of the dataset items above, most are manually created. Exceptions to this are items from the English D2 and English D3 datasets. For English D2, (Olteţeanu et al., 2017) successfully attempted the computational creation of RAT items and

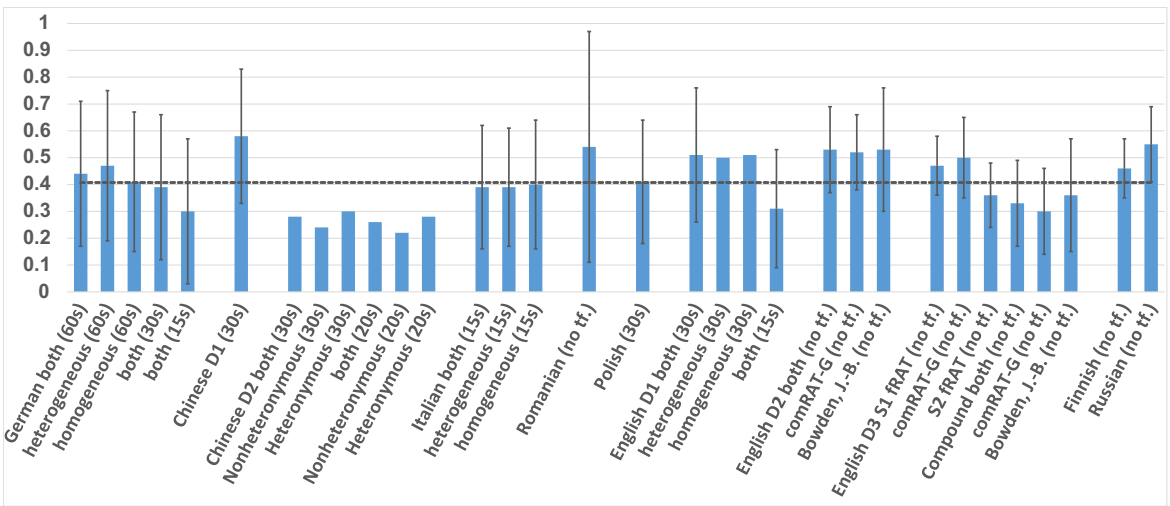

FIGURE 1 | Mean and SD Accuracy of RAT datasets in the different languages.

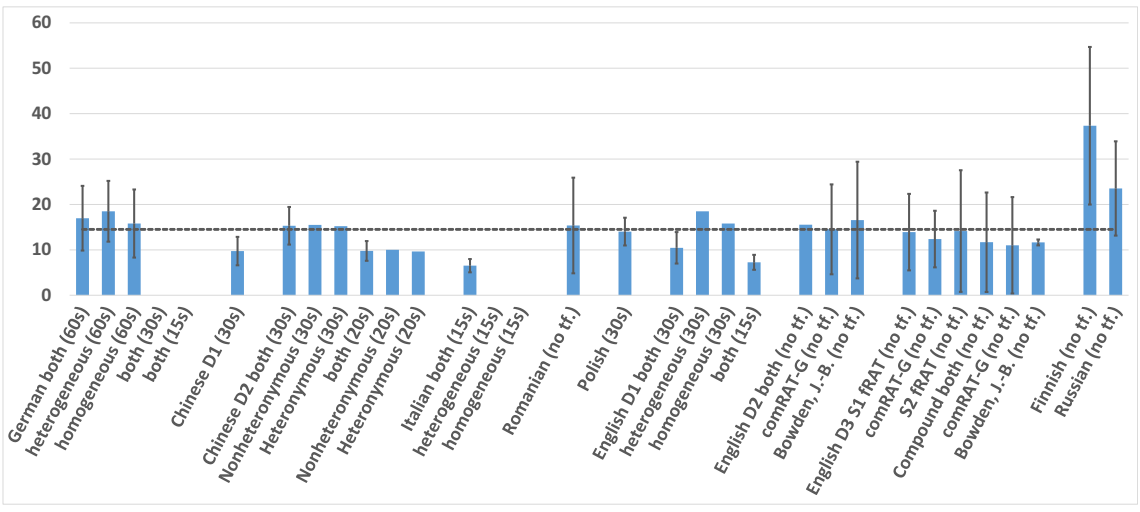

FIGURE 2 | Mean and SD Response Time (in seconds) of RAT datasets in the different languages.

TABLE 1 | 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 different languages.

	Time Frame			Accuracy				RT [s]		Cron- Bach's α	
				Sum		%		Per item			
Language	in s	|x|	n	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	Acc.	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	18.50	6.70	—	—
homogeneous	60	74	80	30.19	19.17	41	26	15.80	7.50	—	—
German both	30	130	80	—	—	39	27	—	—	—	—
German both	15	130	80	—	—	30	27	—	—	—	—
Chinese D1	30	128	123	74.46	—	58	25	9.74	3.13	0.92	—
Chinese D2 both	30	90	71	25.26	—	28	—	15.31	4.14	—	—
Non-heteronymous	30	60	71	18.07	—	24	—	15.49	—	—	—
Heteronymous	30	30	71	7.19	—	30	—	15.21	—	—	—
Chinese D2 both	20	90	93	23.45	—	26	—	9.77	2.17	—	—
Non-heteronymous	20	60	93	16.76	—	22	—	10.01	—	—	—
Heteronymous	20	30	93	6.69	—	28	—	9.65	—	—	—
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	0.93	0.97
Polish	30	17	206	6.90	3.90	41	23	14.02	3.06	0.79	—
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	—	—	0.94	0.99
comRAT-G	None	50	113	26.20	7.03	52	14	14.52	9.89	0.85	0.99
Bowden, J.-B.	None	50	113	26.41	11.24	53	23	16.56	12.84	0.93	0.99
English D3 S1 fRAT	None	75	26	35.27	7.99	47	11	13.91	8.42	—	—
comRAT-G	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	0.79	0.90
Compound both	None	48	61	15.85	7.60	33	16	11.68	10.96	0.87	0.96
comRAT-G	None	24	61	7.25	3.72	30	16	11.00	10.62	0.75	0.93
Bowden, J.-B.	None	24	61	8.61	5.06	36	21	11.64	0.65	0.85	0.92
Finnish	None	47	67	21.60	5.30	46	11	37.34	17.36	0.73	—
Russian	None	48	67	26.60	6.90	55	14	23.53	10.38	0.83	—

S1 and S2 reflect different studies of the same article.

compared the results with an existing (English D1) normative dataset. For English D3, (Olteţeanu et al., 2019a) applied a computational approach using a new type of language knowledge for the creation of functional items, thus resurrecting an older idea of Worthen and Clark (1971) regarding the existence of such items and their differences from compound items. These items are compared to compound items of a subset of English D1 in the paper. This subset—specifically 24 items from English D1—is marked as Bowden, J.-B. in **Figures 1**, **2**.

3.2. Quantitative Comparison

In the following, a descriptive statistics overview of the different datasets is provided.

3.2.1. Descriptive Data

The various RAT datasets contained varying numbers of items, between 17 (Polish) and 144 (English D1). An exception is comRAT-G, which computationally creates 13.4 m items and the frequency-based probabilities of solving them. Furthermore, the various items were deployed either (a) by giving participants different time frames to solve each query, between 2 and 60 s or (b) without setting a time limit. Since 2, 5, 7, 20, and 60-s time frames were only used once across these datasets, only items with a 15 or 30-s time frame or no time frame are analyzed in this paper. Assuming that different solving strategies may be deployed for different time frames, we did not want to average across time frames. The stimuli were

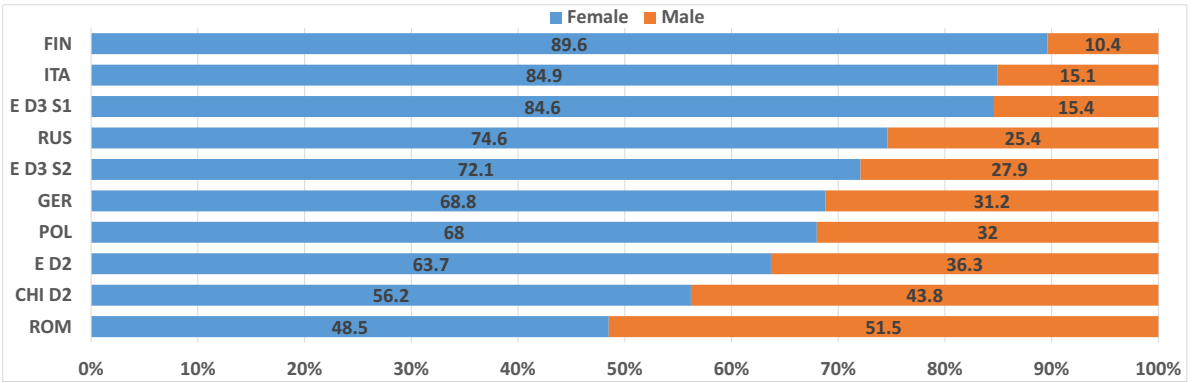

FIGURE 3 | Gender ratio of RAT datasets in the different languages.

TABLE 2 | Results of Mann-Whitney testing with Bonferroni-Holm correction regarding the Accuracy.

Time Frame	Dataset pair		Holm's method			
			p	Rank	α	Sig
15 s	Italian	German	0.0006	3	0.0167	Yes
	Italian	English D1	0.0013	2	0.025	Yes
	German	English D1	0.4694	1	0.05	No
30 s	Chinese D1	Chinese D2	<.0001	21	0.0024	Yes
	Chinese D1	Chinese D2 n.h.	<.0001	20	0.0025	Yes
	Chinese D1	Chinese D2 het.	<.0001	19	0.0026	Yes
	Chinese D2	English D1	<.0001	18	0.0028	Yes
	Chinese D1	German	<.0001	17	0.0029	Yes
	Chinese D2 het.	English D1	<.0001	16	0.0031	Yes
	Chinese D2 n.h.	English D1	<.0001	15	0.0033	Yes
	Chinese D2 het.	Polish	0.0003	14	0.0036	Yes
	Chinese D1	English D1	0.0006	13	0.0038	Yes
	Chinese D2	Polish	0.0006	12	0.0042	Yes
	Chinese D1	Polish	0.0035	11	0.0045	Yes
	Chinese D2 n.h.	Polish	0.0037	10	0.005	Yes
	English D1	German	0.0037	9	0.0056	Yes
	Chinese D2	German	0.0141	8	0.0063	No
	Chinese D2 het.	German	0.0183	7	0.0071	No
	Chinese D2 n.h.	German	0.0889	6	0.0083	No
	English D1	Polish	0.2498	5	0.01	No
	Chinese D2 het.	Chinese D2 n.h.	0.3206	4	0.0125	No
	German	Polish	0.4048	3	0.0167	No
	Chinese D2	Chinese D2 het.	0.4822	2	0.025	No
	Chinese D2	Chinese D2 n.h.	0.6559	1	0.05	No
None	English D3 S2 fRAT	Romanian	<0.0001	10	0.005	Yes
	English D3 S2 fRAT	Russian	0.0006	9	0.0056	Yes
	English D2	English D3 S2 fRAT	0.0054	8	0.0063	Yes
	English D3 S2 fRAT	Finnish	0.0301	7	0.0071	No
	Finnish	Russian	0.0698	6	0.0083	No
	Finnish	Romanian	0.0884	5	0.01	No
	English D2	Finnish	0.3515	4	0.0125	No
	English D2	Russian	0.7583	3	0.0167	No
	Romanian	Russian	0.9033	2	0.025	No
	English D2	Romanian	0.9492	1	0.05	No

deployed on populations of various sizes, with n ranging between 26 participants in the English D3 S1 and 317 in the Italian dataset.

As shown in **Table 1**, **Figures 1**, **2**, the easiest sets to solve were the Chinese D1, with 0.58 accuracy, and the Italian, with a response time of only 6.52 s. The hardest sets seem to be the Chinese D2, with an average accuracy of 0.26 within a 20-s time frame, and the Finnish dataset in terms of response times, with a mean of 37.34 s. The response times of the Russian RAT were also noticeably higher that for the rest (23.53 s). 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 authors.

The age, level of education, and gender of the participants taking the different RATs also varied, as shown in **Tables A1–A5**, and **Figure 3**. For example, 70% of the participants of the Russian RAT were between 20 and 29 years old, whereas over 50% of the English D3S2 were between 30 and 39 years old. The Romanian RAT had the most equal gender ratio, at nearly 50/50, while the Finnish had the worst, with 90% females. **Table 1** gives an overview of all of the datasets and various descriptive metrics across all languages.

3.2.2. 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's α scores were calculated by authors for some of the initial papers (see **Table 1**) and vary between 0.73 and 0.99.

4. RESULTS

4.1. Language

In order to find out whether differences between results for different languages exist at all, Kruskal-Wallis Tests were conducted for different timesteps and on two existing performance metrics: Accuracy and Response Time. To further investigate which of the language pairings were different, we used pairwise Mann-Whitney tests with Bonferroni-Holm correction as *post-hoc* tests. Heterogeneous and homogeneous items were tested both separately and combined (where possible).

4.1.1. Accuracy

We found a significant effect of group on value for the 30-s time frame [$\chi^2_{(6)} = 110.05, p < 0.0001$], the 15-s time frame [$\chi^2_{(2)} = 14.58, p < 0.001$], and for no time frame [$\chi^2_{(4)} = 18.36, p < 0.01$]. *Post-hoc* tests showed significant differences of means regarding the Accuracy metric for 18 different dataset pairings in different time frames (**Table 2**). For example, a significant difference exists between Italian vs. German in a 15-s time frame ($p = 0.00062$, $\alpha = 0.01667$).

TABLE 3 | Results of Mann-Whitney testing with Bonferroni-Holm correction regarding the RT.

Time	Dataset pair		Holm's method			
Frame			*p*	Rank	α	Sig
15 s	English D1	Italian	<0.0001	1	0.05	Yes
30 s	Chinese D2	Chinese D1	<0.0001	15	0.0033	Yes
	Chinese D2	English D1	<0.0001	14	0.0036	Yes
	Chinese D2 n.h.	Chinese D1	<0.0001	13	0.0038	Yes
	Chinese D2 n.h.	English D1	<0.0001	12	0.0042	Yes
	Chinese D2 het.	Chinese D1	<0.0001	11	0.0045	Yes
	Chinese D2 het.	English D1	<0.0001	10	0.005	Yes
	Chinese D1	Polish	<0.0001	9	0.0056	Yes
	English D1	Polish	<0.0001	8	0.0063	Yes
	Chinese D1	English D1	0.1201	7	0.0071	No
	Chinese D2 n.h.	Polish	0.2384	6	0.0083	No
	Chinese D2	Polish	0.2838	5	0.01	No
	Chinese D2 het.	Polish	0.5176	4	0.0125	No
	Chinese D2 het.	Chinese D2 n.h.	0.9018	3	0.0167	No
	Chinese D2	Chinese D2 het.	0.9304	2	0.025	No
	Chinese D2	Chinese D2 n.h.	0.9558	1	0.05	No
None	Finnish	Romanian	<0.0001	10	0.005	Yes
	Finnish	English D3 S2 fRAT	<0.0001	9	0.0056	Yes
	Finnish	English D2	<0.0001	8	0.0063	Yes
	Russian	English D3 S2 fRAT	<0.0001	7	0.0071	Yes
	Russian	Romanian	<0.0001	6	0.0083	Yes
	Finnish	Russian	<0.0001	5	0.01	Yes
	English D2	English D3 S2 fRAT	0.0095	4	0.0125	Yes
	English D3 S2 fRAT	Romanian	0.0701	3	0.0167	No
	English D2	Romanian	0.0749	2	0.025	No
	English D2	Russian	0.1370	1	0.05	No

4.1.2. Response Time

We found a significant effect of group on value for the 30-s time frame [$\chi^2_{(5)} = 158.76, p < 0.0001$] and for no time frame [$\chi^2_{(4)} = 64.74, p < 0.0001$]. *Post-hoc* tests using Mann-Whitney tests with Bonferroni-Holm correction showed significant differences of means regarding the Response Time metric for 16 different dataset pairings in different time frames (**Table 3**). For example, a significant difference was noted between English D2 vs. English D3 S2 fRAT with no time frame ($p = 0.0095, \alpha = 0.0125$).

4.2. Gender

In order to measure differences between genders, Welch's unequal variances t-test was conducted to measure the difference between means on the two existing performance metrics: Accuracy and Response Time. Moreover, Cohen's d was calculated to measure the effect size.

4.2.1. Differences Between Genders

Significant differences of means for Accuracy with medium effect sizes were observed between genders in:

(i) Romanian; $t(59.47) = 2.29^*$, male $M = 55.25$, female $M = 64.61$
(ii) English D3; $t(24.17) = 2.21^*$, male $M = 29.78$, female $M = 37.88$

as shown in **Table A5**, but no differences were observed regarding the Response Time. The authors of the Chinese D2 and an older

Italian version (Salvi et al., 2015) also stated that gender was not a factor in their experiments.

5. DISCUSSION AND FURTHER WORK

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 Time 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 vs. functional).

In the cases in which a significant difference exists between different language datasets, the main potential causes are:

(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 and thus depend on the inspiration and knowledge base of the researchers creating them, or
(d) the lack of a common time frame.

Other causes could be, as pointed out by our reviewers, differences in the instructions/explanation of the task, in participants' motivations, in the study setting (e.g., fMRI scanner/EEG, etc.), in other tasks performed during the same session, and in whether solution feedback was given, and, also, the associations between the items themselves might affect the difficulty (Luft et al., 2018).

This initial investigation shows that differences between results obtained with the RAT in different languages need to be addressed in more detail. Before cross-comparison of creativity results can be performed, the source of these differences needs to be found. Experimental or analytical setups need to be designed in order to establish which one of the above-mentioned causes, or what combination thereof, is the source of the differences.

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

However, even if translatable, the same RAT items may not be of the same difficulty in different languages. Some light is shed on this by computational models like comRAT-C (Olteţeanu and Falomir, 2015), essentially models of memory search, which can solve the RAT by organizing their knowledge in a semantic net-like structure, propagating activation through word associations and convergence. comRAT-C's probability of solving a query correlates with human performance. Such models indicate 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 and Schultheis, 2017) and in the creation of other types of items (Olteţeanu et al., 2019a). We have not yet attempted to generate comparable RAT stimulus sets in different languages. To apply the computational approach above for RAT generation in multiple languages would require initial sets of word associations or n-grams for each of the respective languages together with data on how often the n-grams occur within a specific dataset or the frequency of eliciting a particular associate (if a certain number of participants is asked to produce associates).

Another direction of future work would be to establish a creative association measure that transcends the constraints of language such as a visual Remote Associates Test—some work in this direction has already been done by Olteţeanu et al. (2015) and Toivainen et al. (2019). As one of our reviewers very interestingly points out, visual information, though not as varied as language, nonetheless varies in different cultures. The visual RATs would thus not be completely immune to differences, for example, when apple trees are more common in some parts of the world and mango trees in others (our reviewer's example) or when certain objects are more likely to exist, be used, or be central in various cultures. However, these difference may be smaller than linguistic differences for specific sets of objects, and the visual RAT may thus provide a measure with stronger comparability across languages.

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.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

A-MO contributed the conception and design of the study and wrote the first draft of the manuscript. JB performed the statistical analysis. JB and A-MO wrote sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

The support of the Deutsche Forschungsgemeinschaft (DFG) for the project CreaCogs via grant OL 518/1-1 and the Open Access Funding provided by the Freie Universität Berlin are gratefully acknowledged.

REFERENCES

Ansburg, P. I., and Hill, K. (2003). Creative and analytic thinkers differ in their use of attentional resources. *Pers. Individ. Diff.* 34, 1141–1152. doi: 10.1016/S0191-8869(02)00104-6

Baba, Y. (1982). JARAT FORM A-remote associates test. *Jpn. J. Psychol.* 52, 330–336. doi: 10.4992/jjpsy.52.330

Bowden, E. M., and Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behav. Res. Methods* 35, 634–639. doi: 10.3758/BF03195543

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., and Mednick, S. C. (2009). Rem, not incubation, improves creativity by priming associative networks. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10130–10134. doi: 10.1073/pnas.0900271106

Chermahini, S. A., Hickendorff, M., and Hommel, B. (2012). Development and validity of a Dutch version of the Remote Associates Task: an item-response theory approach. *Think. Skills Creat.* 7, 177–186. doi: 10.1016/j.tsc.2012.02.003

Cunningham, J. B., MacGregor, J., Gibb, J., and 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. *J. Creat. Behav.* 43, 262–280. doi: 10.1002/j.2162-6057.2009.tb01318.x

Druzhinin, V. N. (1999). *Psychology of General Abilities*. St. Petersburg: Publishing House Peter.

Jen, C.-H., Chen, H.-C., Lien, and Cho (2004). The development of the Chinese remote association test. *Res. Appl. Psychol.* 21, 195–217.

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

Luft, C. D. B., Zioga, I., Thompson, N. M., Banissy, M. J., and Bhattacharya, J. (2018). Right temporal alpha oscillations as a neural mechanism for inhibiting obvious associations. *Proc. Natl. Acad. Sci. U.S.A.* 115, E12144–E12152. doi: 10.1073/pnas.1811465115

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

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

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

Olteţeanu, A.-M., Gautam, B., and 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 (Turin: CEUR-Ws), 19–28.

Olteţeanu, A.-M., Schoettner, M., and Schuberth, S. (2019a). Computationally resurrecting the functional Remote Associates Test using cognitive word associates and principles from a computational solver. *Knowl. Based Syst.* 168, 1–9. doi: 10.1016/j.knosys.2018.12.023

Olteţeanu, A.-M., and Schultheis, H. (2017). What determines creative association? revealing two factors which separately influence the creative process when solving the remote associates test. *J. Creat. Behav.* 53, 389–395. doi: 10.1002/jocb.177

Olteţeanu, A.-M., Schultheis, H., and Dyer, J. B. (2017). Computationally constructing a repository of compound Remote Associates Test items in American English with comRAT-G. *Behav. Res. Methods* 50, 1971–1980. doi: 10.3758/s13428-017-0965-8

Olteţeanu, A.-M., Taranu, M., and Ionescu, T. (2019b). Normative data for 111 compound Remote Associates Test problems in Romanian. *Front. Psychol.* 10:1859. doi: 10.3389/fpsyg.2019.01859

Orita, R., Hattori, M., and Nishida, Y. (2018). Development of a Japanese remote associates task as insight problems. *Jpn. J. Psychol.* 89, 376–386. doi: 10.4992/jjpsy.89.17201

Salvi, C., Bricolo, E., Franconeri, S. L., Kounios, J., and Beeman, M. (2015). Sudden insight is associated with shutting out visual inputs. *Psychon. Bull. Rev.* 22, 1814–1819. doi: 10.3758/s13423-015-0845-0

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

Shen, W., Yuan, Y., Liu, C., Yi, B., and Dou, K. (2016). The development and validity of a Chinese version of the compound remote associates test. *Am. J. Psychol.* 129, 245–258. doi: 10.5406/amerjpsyc.129.3.0245

Sio, U. N., and Rudowicz, E. (2007). The role of an incubation period in creative problem solving. *Creat. Res. J.* 19, 307–318. doi: 10.1080/10400410701397453

Sobków, A., Połeć, A., and Nosal, C. (2016). Rat-pl- construction and validation of polish version of remote associates test. *Stud. Psychol.* 54, 1–13. doi: 10.2478/V1067-010-0152-2

Toivainen, T., Olteţeanu, A.-M., Repykova, V., Lihanov, M., and Kovas, Y. (2019). Visual and linguistic stimuli in the Remote Associates Test: a cross-cultural investigation. *Front. Psychol.* 10:926. doi: 10.3389/fpsyg.2019.00926

Ward, J., Thompson-Lake, D., Ely, R., and Kaminski, F. (2008). Synaesthesia, creativity and art: what is the link? *Br. J. Psychol.* 99, 127–141. doi: 10.1348/000712607X204164

Worthen, B. R., and Clark, P. M. (1971). Toward an improved measure of remote associational ability. *J. Educ. Meas.* 8, 113–123. doi: 10.1111/j.1745-3984.1971.tb00914.x

Wu, C.-L., and Chen, H.-C. (2017). Normative data for Chinese compound remote associate problems. *Behav. Res. Methods* 49, 2163–2172. doi: 10.3758/s13428-016-0849-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

TABLE A1 | Mean, standard deviation, and range of participant age.

	ER	CHI_{D2}	ITA	POL	ROM
M_{age}	23.41	22.67	25.30	25.10	—
SD_{age}	2.93	3.24	8.30	7.60	—
$range_{age}$	20–31	18–34	16–65	18–55	18–70

TABLE A2 | Percentage of participants in certain age ranges.

	ROM	E_{D2}	E_{D3S1}	E_{D3S2}	FIN	RUS
<20	4.8	1.8	0.0	2.0	0.0	4.5
20–29	42.9	29.2	11.5	23.0	22.4	70.1
30–39	30.1	31.9	53.8	21.0	23.9	22.4
40–49	14.3	15.0	11.5	20.0	28.3	1.5
50–59	6.3	17.7	23.1	25.0	20.9	1.5
59<	1.6	4.4	0.0	10.0	4.5	0.0

TABLE A3 | Percentage of participants with a certain level of education.

	ROM	E_{D2}	E_{D3S1}	E_{D3S2}
Secondary school	0.0	6.2	11.5	5.0
High school diploma	14.3	23.9	26.9	25.0
Enrolled in undergraduate courses	17.5	17.7	15.4	7.0
Completed undergraduate courses	50.8	30.1	23.1	52.0
Enrolled in postgraduate courses	4.8	5.3	11.5	2.0
Completed postgraduate courses	12.7	16.8	11.5	10.0

TABLE A4 | Percentage of participants with certain gender.

	GER	CHI_{D2}	ITA	POL	ROM	E_{D2}	E_{D3S1}	E_{D3S2}	FIN	RUS
Female	68.8	56.2	84.9	68.0	48.5	63.7	84.6	72.1	89.6	74.6
Male	31.2	43.8	15.1	32.0	51.5	36.3	15.4	27.9	10.4	25.4

TABLE A5 | Welch test results for accuracy without a time frame between genders.

	ROM female				ENG D3 S2 fRAT female			
	t	df	*p*	d	*t*	df	*p*	d
ROM male	2.29	59.47	0.03	0.58	—	—	—	—
ENG D3 S2 fRAT male	—	—	—	—	2.21	24.17	0.04	0.70