Modeling the Remote Associates Test as Retrievals from Semantic Memory

March 25th, 2022

Jule Schatz (schatzju@umich.edu)
Steven J. Jones (scijones@umich.edu)
John E. Laird (laird@umich.edu)
University of Michigan, 2260 Hayward Street
Ann Arbor, MI 48109-2121 USA

Keywords: Cognitive Modeling; Semantic Memory; Remote Associates Test; Cognitive Architecture

Modeling the Remote Associates Test as Retrievals from Semantic Memory

March 25, 2022

Abstract

The Remote Associates Test (RAT) is a word association retrieval task, which consists of a series of problems, each with three seemingly unrelated prompt words. The subject is asked to produce a single word that is related to all three prompt words. In this paper, we provide support for a theory in which the RAT assesses a person's ability to retrieve relevant word associations from long-term memory. We present a computational model of humans solving the RAT and investigate how prior knowledge and memory retrieval mechanisms influence the model's ability to match human behavior. We expand prior modeling attempts by investigating multiple large knowledge bases and by creating a cognitive process model that uses long-term memory spreading activation retrieval processes inspired by ACT-R and implemented in Soar. We evaluate multiple model variants for their ability to model human problem difficulty, including the incorporation of noise and base-level activation into memory retrieval. We conclude that the main factors affecting human difficulty are: the existence of associations between prompt words and solutions, the relative strengths and directions of those associations compared to associations to other words, and the ability to perform multiple retrievals.

1 Introduction

The Remote Associates Test (RAT) is a series of problems, where each problem consists of presenting the subject with three *prompt* words. The subject must determine a fourth *solution* word that is associated with all three prompt words. For example, if "swiss," "cake," and "cottage" are the prompt words, then "cheese" is the sought-after response. The work reported here, as well as in multiple past research projects (Wu, Huang, Chen, & Chen, 2020), uses 144 compound RAT problems and the associated human performance on those problems (Bowden & Jung-Beeman, 2003). The original RAT test was developed by Mednick (1962).

In this paper, we provide support for a theory where performance on the RAT is determined by knowledge in long-term declarative memory and iterative retrieval using well-established retrieval mechanisms that rely on activation of words in memory and their connections. The model of activation we use consists of Base-Level Activation (BLA), spreading activation, and noise (Anderson, 1990).

$$Activation = BaseLevelActivation + SpreadingActivation + Noise \tag{1}$$

To explore this theory, we develop a family of models of RAT performance using different knowledge bases, differing numbers of attempted retrievals, and declarative memory retrieval mechanisms inspired by ACT-R as implemented in Soar (Jones, Wandzel, & Laird, 2016). The base model relies on spreading activation to govern retrievals. However, we also explore variants that incorporate BLA and different levels of noise.

Prior research has created and evaluated computational models based on retrieval from long-term memory (described in the Related Work section); however, our work is distinguished by our comprehensive explorations and evaluations of a wide range of variants that allow us to determine the contributions of the underlying knowledge base, word associations, the directions and strengths of those associations, BLA, noise, and the number of retrieval attempts. Our explorations allow us to disentangle potential interactions among these factors, helping us create a detailed model of semantic memory retrieval for this task.

In the remainder of the paper, we first examine the RAT. Next, we discuss five knowledge bases that fall into two major categories and analyze them as to their suitability for modeling the RAT. The two categories of knowledge bases include ones built from word co-occurrences in text (COCA-TG, Google books) and ones built from word associations collected from humans (HBC, USF Norms, SWOWEN).

We then present a simple computational process model with variants along the dimensions described earlier. We evaluate the models on how well they match human performance on the RAT in terms of *problem difficulty* for two conditions: one where humans have 7 seconds to generate an answer and one where they have 15 seconds. We also compare the computational process models to Distributional Semantic Models (DSMs).

The best model uses human generated associations (SWOWEN) and includes spreading activation attenuated by association strength and noise to retrieve words from long-term memory, employing multiple attempts until it finds a potential answer or exhausts its number of attempts. It models human question difficulty on the RAT with a R^2 of 0.90 and a mean-squared error of 0.63 on the 7 second task and with a R^2 of 0.98 and a mean-squared error of 0.29 on the 15 second task. The only variation between modeling the 7 second task and the 15 second task is the number of retrieval attempts allowed to find a solution. In matching human performance, there is a linear trade off between the amount of noise and the number of attempts. For a given amount of time, the main determiners of difficulty of the RAT problems are directionality and strength of the associations between words.

Table 1
Remote Associate Test example problems.

Prompt Word 1	Prompt Word 2	Prompt Word 3	Solution Word
man	glue	star	super
dew	comb	bee	honey
rain	test	stomach	acid

2 Remote Associates Test

Bowden and Jung-Beeman (2003) developed the 144 compound RAT problems used in this work. Three examples are shown in Table 1. The RAT problems are designed to have only a single valid answer. To minimize any variance from confounding factors, Bowden and Jung-Beeman (2003) used only compound words or phrase associations when creating the RAT items. For example, the association between "super" and "man" is valid because those two words are found next to each other in the English language. The association between "uncle" and "man" is not valid, even though they are semantically related, because "uncle" and "man" do not form a compound word or common phrase.

All RAT problems use common words that are intended to be familiar to the participants. To avoid priming interference effects, solution words are never repeated nor used as prompt words. The overall test is difficult for humans, who on average correctly answer 32.92 out of 144 compound RAT questions given 7 seconds and 44.25 given 15 seconds. Although generating a correct response is difficult, once the correct response is provided, most subjects recognize it as an appropriate answer. This suggests that task difficulty is not because the necessary knowledge is missing, but because it is difficult to access.

The associations between prompt words and the solution word vary in direction. For example, with the RAT problem "rain," "test," and "stomach," the direction of some associations go from a prompt word to the solution word (such as from "stomach" to "acid"), which we call *forward associations*. Others go from the solution word to a prompt word (such as from "acid" to "rain"), which we call *backward associations*. After directly inspecting each problem, we observe that each word pair has a single correct direction that creates a compound word or common phrase. In 30.3% of all word pairs, the correct association direction is forward, whereas in the remaining the direction is backward. While not discussed by Bowden and Jung-Beeman (2003) or other related work as being an important aspect of individual problems, we observe that questions that have more forward associations tend to be easier for humans to solve than questions that have more backward associations, with a correlation of 0.87.

Bowden and Jung-Beeman (2003) provided data from a human study where participants solve the 144 compound RAT problems. We use these data in our evaluation of the knowledge bases and models. It provides data on 289 participants that were given either 2, 7, 15, or 30 seconds to produce a solution for each of the 144 compound RAT problems. We do not model the 30 second data because those experiments took significantly longer (up to 4 hours per full test) with the subjects in an fMRI machine. We do not include the 2 second data in our initial development and evaluation of our models, in part, because on average, humans get so few of those questions correct (8.0 out of 144). We do summarize the models' performance on those data at the beginning of the Discussion and Analysis section. Thus, we focus the development and evaluation of our models on the 7 and 15 second data. The correlation for percentage of participants producing solutions within 7 seconds and 15 seconds is 0.91, meaning that problems that were difficult in 7 seconds also tended to be difficult in 15 seconds.

3 Knowledge Bases

A consistent hypothesis across computational models of the RAT is that performance is strongly dependent on the contents of its knowledge base, including the associations between words, and the direction and strength of those associations.One of the contributions of this work is evaluating five different word association knowledge bases that come from two general categories: word co-occurrences and collected word associations. In this work, as in many of the models described in the Related Work section, the knowledge base is a weighted directed graph. The nodes represent words while the edges between the words represent associations between words. Edges are directed, unlabeled, and weighted, where the weight represents the strength of the association and the direction represents the direction of the association.

Word co-occurrence models derive associations from words that co-occur next to each other in text. We explored two models, one derived from the Corpus of Contemporary American English (COCA) (Davies, 2008) and one derived from Google Books (Lin et al., 2012).

Olteţeanu and Falomir (2015) created a word co-occurrence knowledge base based on COCA to solve the RAT called RAT-KB. We attempted to recreate their knowledge base by extracting the top 2-grams from COCA and processed them as described in their paper. However, our replicated knowledge base does not match their knowledge base, even when following their exact instructions.[1] We expanded RAT-KB to increase its coverage of RAT problems by allowing all parts of speech instead of the limited hand selected set that RAT-KB used, a version we call COCA-TG (COCA Two-Grams). All associations in COCA-TG are bidirectional and include strengths based on their frequency in COCA. Since COCA-TG uses only the top 2-grams in COCA, it does not include semantic or relational associations, such as a relationship between "man" and "uncle." COCA-TG has 55,999 unique words, 852,217 association links, and 100% of the RAT problem words. In 94 out of the 144 RAT problems, COCA-TG has associations between all three prompt words and the solution word. However, getting these and other RAT problems correct still depends on the algorithm used with the knowledge base.

[1] We contacted the authors, but they were unable to provide insight into this difference.

Google Books is a corpus derived from an analysis of over five million books. This knowledge base has been used in previous RAT modeling research (Kajić, Gosmann, Stewart, Wennekers, & Eliasmith, 2016, 2017). Similar to COCA-TG, this knowledge base is built from word co-occurrences and consists of 2-grams and their frequency in English. Word co-occurrences for Google Books are directional. If word1 appears in the texts before word2, it is collected as an association from word1 to word2 and not vice versa. It has about 90 times more word associations and 28 times more unique words than COCA-TG. Because it contains data on the co-occurrence of essentially all English words with each other, it is not biased toward the compound word phrases (such as “cottage cheese”), and many words have their highest associations to function words (also called “stop” words), such as “and,” “the,” or “of,” making it problematic as a source for associational retrievals of solution words for the RAT. To avoid these issues, we removed all stop words from the knowledge base. After the processing, the knowledge base has 100% of the words related to the 144 RAT items and full solutions to all 144 RAT problems.[2]

In contrast to using large corpora of written English, the other databases were built by collecting word associations directly from human participants through a study and/or game. These games present a participant with a cue word and then ask for the participant to respond with the first word they think of that is associated with the given word. For example, if the given word is “bird,” a participant might respond with “feather,” “brain,” or “fly.” Data are collected in the form of “word1,” and “word2,” with repeated occurrences summed to give the number of instances “word2” was entered when a participant saw “word1.” Thus, there is a one-way *association* or *link* created from “word1” to “word2,” and the sum is used as the *association strength*. We investigate three large databases that collected word associations in this way: Human Brain Cloud (HBC) (Gabler, 2013), University of South Florida Free Association Norms (USF Norms) (Nelson, McEvoy, & Schreiber, 1998), and Small WOrld of Words in ENglish (SWOWEN) (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2018).

HBC is crowd-sourced through an online game of word associations, where a player is presented with an eligible word chosen at random and then asked to type another word that they believe to be closely related to the given word. The website started with a single eligible word, “volcano.” All additional words become eligible to be presented when they have been typed in at least 20 times as a related word. After removing items with multiple words and non-alphabetic characters, there are 231,858 unique words and 2,403,813 associations between those words. This data set has 100% of the words in the 144 RAT items and full solutions to 109 out of 144 of the RAT problems. This knowledge base has not been used for modeling the RAT in previous work.

USF Norms is the accumulation of data collected using multiple methodologies. After removing items with multiple words and non-alphabetic characters, there are 9,883 unique words and 70,699

[2] By full solutions we mean that there exist connections in at least one direction between all three prompt words and the solution word for every RAT problem.

associations between those words. This data set has 97.7% of the words related to the 144 RAT items and full solutions to 21 out of 144 of the RAT problems. This knowledge base has been used in previous RAT modeling research (Kajić et al., 2016, 2017).

SWOWEN data was built from a demographically diverse group of participants. Cue words started as a limited set from previous research on semantic categories. New cue words were added when they were frequently stated as responses, similar to HBC. After removing items with multiple words and non-alphabetic characters, there are 38,339 unique words, 460,938 associations between those words. This data set has 100% of the words related to the 144 RAT items and full solutions to 51 out of 144 of the RAT problems. This knowledge based was used in previous work modeling the RAT (Valba, Gorsky, Nechaev, & Tamm, 2021).

Table 2 summarizes the statistics of the five knowledge bases. Google Books contains the most unique words, the most associations, all of the related RAT words, and all the solution links. The next biggest knowledge base is HBC, which is the largest one collected directly as word associations. Further analysis of the knowledge bases in terms of modeling performance on the RAT is discussed in the Knowledge Base Baselines section and the Model Experiments and Evaluation section.

4 Process Model of the RAT

Our hypothesis is that the ACT-R model of retrieval from long-term memory (reimplemented in Soar), can successfully model human performance on the RAT. Our model focuses on memory retrieval and internal task processing but does not include initial perception and comprehension (reading) of the prompt words, eye movements and re-perception of words, or typing a response. Furthermore, our model does not include complex strategies, such as focusing on a single word or pair of words during retrieval. These strategies could influence performance; however, our goal here is to develop and evaluate a simple model based on well established theories of retrieval from long-term memory. Thus, our model does not attempt to account for individual differences that arise from variation in these and other strategies, nor does it attempt to accurately model the detailed time course of producing an answer. As further described in the Experiments and Evaluation section, the purpose is to model

Table 2
Knowledge base summary.

	COCA-TG	Google Books	HBC	USF	SWOWEN
Unique Words	55,999	1,678,975	231,858	9,883	38,339
% RAT Words	100%	100%	100%	97.7%	100%
Assoc. Links	852,217	72,843,006	2,403,813	70,699	460,938
Solution Links	94	144	109	21	51

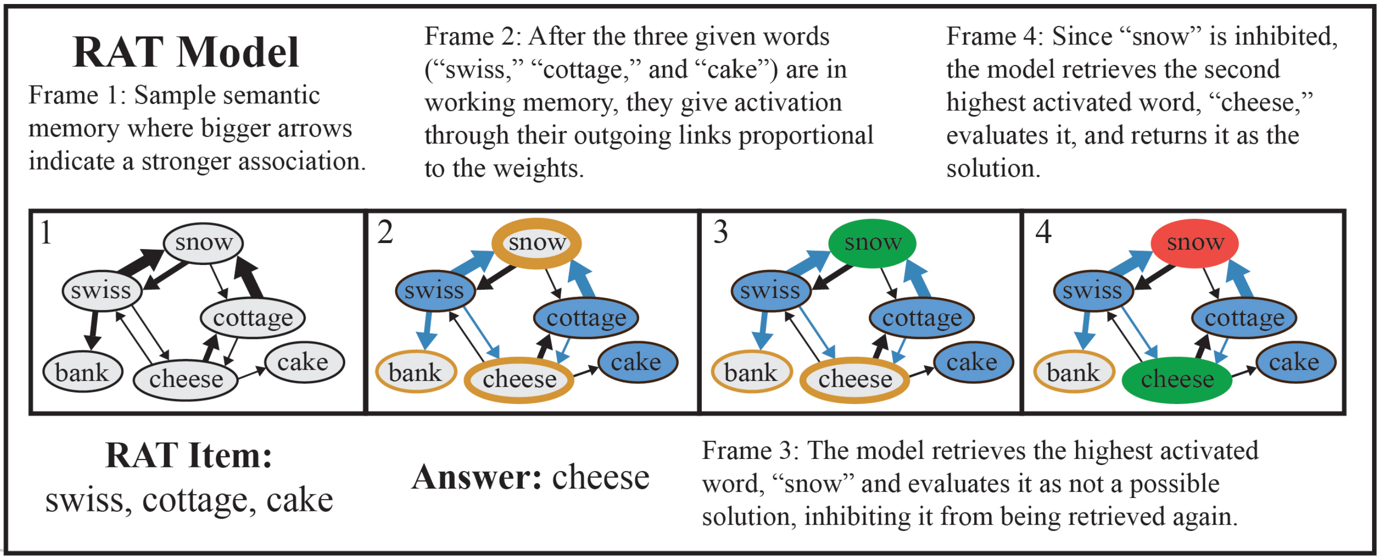
RAT Model
Frame 1: Sample semantic memory where bigger arrows indicate a stronger association.
Frame 2: After the three given words ("swiss," "cottage," and "cake") are in working memory, they give activation through their outgoing links proportional to the weights.
Frame 4: Since "snow" is inhibited, the model retrieves the second highest activated word, "cheese," evaluates it, and returns it as the solution.
1
snow
swiss
cottage
bank
cheese
cake
2
snow
swiss
cottage
bank
cheese
cake
3
snow
swiss
cottage
bank
cheese
cake
4
snow
swiss
cottage
bank
cheese
cake
RAT Item:
swiss, cottage, cake
Answer: cheese
Frame 3: The model retrieves the highest activated word, "snow" and evaluates it as not a possible solution, inhibiting it from being retrieved again.

Fig. 1. A walk through of a RAT example using the model.

problem difficulty averaged across a population of subjects.

In our model, all words and their associations are stored in long-term declarative memory, which in Soar is called semantic memory. Furthermore, words are retrieved using ACT-R's declarative memory mechanisms (described below) as implemented in Soar (Derbinsky & Laird, 2010; Jones et al., 2016). Given that our model replicates the declarative memory retrieval mechanisms in ACT-R, we expect that results from an ACT-R model would be similar. However, given the variability in how a model and the accompanying knowledge base can be encoded in a cognitive architecture, we refrain from making any definitive claims about potential ACT-R models.

Our base model is a straightforward implementation of repeated retrievals from long-term memory biased by spreading activation. In our model, semantic memory is initialized with the word associations from a knowledge base, where words are represented as nodes and associations are represented as directed weighted edges. The model receives the spelling of the three prompt words in working memory. To comprehend each word, the model uses the word's spelling to retrieve it from semantic memory into working memory. The presence of these words in working memory causes the corresponding words in semantic memory to activate and spread activation along outgoing forward links to other words. The algorithm for spreading activation uses only forward associations to spread activation, which in turn directly determines activation of words in semantic memory. Frame 1 of Figure 1, shows a sample semantic memory where connections with larger arrows indicate stronger associations. In frame 2, the model retrieves the three prompt words, "swiss," "cottage," and "cake" into working memory, which causes the words to activate and spread activation along outgoing forward links in proportion to the association strength of each link. After the prompt words have been retrieved, the model attempts to retrieve the solution word from semantic memory, biased by spreading activation.

The spreading activation equation used by our model is given below:

$$S_j = \log\left[\sum_i \left(\frac{a_{ij}}{fan_i}\right)\right] + offset \tag{2}$$

where the sources of spreading are i, a recipient of spreading is j, an association weight from i to j is a_{ij}, and fan_i is the outgoing fan for source i. The *offset* provides a minimum value for spread. Spreading is limited to a depth of one because the task involves compound words or phrase associations that are presumably only one link away. Further details on our semantic memory parameters and spreading activation implementation can be found in Appendix A. In some models of spread, there is a separate fan term that attenuates the spread based on the number of outgoing links from a source word. For some knowledge bases, the number of distinct responses from a source word could be interpreted as fan. However, this apparent fan depends on an how often a source word was presented to users during data collection instead of the number of associations an average user has from a word. This makes the apparent fan biased because different source words were presented with different frequencies. Thus, we normalize the outgoing association strengths from a source word. We use these normalized association strengths as a_{ij}. This normalization avoids bias from differing word presentation frequencies and limits the associations to sum to 1. We then set fan_i to 1 so not to reintroduce the bias from data collection.

As shown in frame 3, the first attempt retrieves the highest activated word, in this case "snow." High activation by itself is no guarantee that the retrieved word is associated with all three prompt words, as a high activation strength from one or two prompt words can dominate. This is the case in our example, where there are links from "swiss" and "cottage" to "snow," but no link from "cake." The model tests to see if a retrieved word is a solution by testing if it has a link in either direction with every prompt word. Thus, the algorithm for evaluating potential solutions to RAT problems uses both forward and backward associations from semantic memory, although retrieval uses only forward links. If the retrieved word has at least one link with each prompt word, the model uses that word as the solution. If not, as in this case where "snow" is not linked to "cake," the model queries semantic memory again, inhibiting any words it has previously retrieved, which in our example is "snow," as shown in red. In Soar, inhibiting a word from being retrieved is a simple specification to the retrieval link done through procedural memory rules. In ACT-R, inhibition is implemented through a transient decrease to the activation of a retrieved memory element (Lebiere & Best, 2009). In frame 4, the model retrieves the second highest activated word "cheese," which is associated with all three prompt words, so the model returns it as the solution. The maximum number of retrievals for a solution is a parameter called *attempts*, which we experimentally determined to best match human performance, as described in the Model Experiments and Evaluation section. If the model exhausts the number of attempts, it "guesses" by returning the retrieved word that has associations with the most prompt words. Ties are broken randomly. To account for the impact of this random behavior, we ran each model 100 times and averaged the results. An alternative model was tried where ties were broken by

Table 3

Iterative processing of the model with HBC for the RAT problem: “dew,” “comb,” and “bee.”

	Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5
	mountain	hair	brush	sting	honey
dew	0.37	0.00	0.00	0.00	0.01
comb	0.00	0.35	0.29	0.00	0.03
bee	0.00	0.00	0.00	0.15	0.10

Note. For each attempt, the association strength from each prompt word is reported.

choosing the retrieved word with the highest activation. This model change had little to no impact on the results.

To illustrate the iterative processing of the model, Table 3 shows the RAT problem where “dew,” “comb,” and “bee” are prompt words and where the solution word is “honey.” The table shows the five attempts that the model, using HBC as its knowledge base, makes as it retrieves words (“mountain,” “hair,” ...) before reaching the solution. For each attempt, the table shows the word with the highest activation on that attempt across the top, and the contribution of activation spread from each prompt word for that word below it. At first, words are selected where there is a strong forward association from a single prompt word. (0.000 indicates no association.) At attempt 5, the combined forward association strengths of all three prompt words provides the highest activation, leading to the retrieval of the correct answer. As shown, the strength of a strong forward association from a single prompt word can dominate the contributions of multiple weak forward associations. However, in this case, it is exactly those weak, but non-zero forward associations that are necessary to retrieve the correct answer.

Previous work that examined human strategies for solving the RAT concluded that people do not hone in on a solution to all three words at once but instead look at one word at a time and search through associated word spaces (Davelaar, 2015; Smith, Huber, & Vul, 2013). Specifically, Smith et al. (2013) found that people report potential solutions that are associated with only one of the prompt words before producing the correct solution. While our model does not explicitly focus on one prompt word at a time, it often follows a pattern of initially retrieving words that are strongly associated with only one of the prompt words. Individual response traces were not available as part the data we analyzed, and further work needs to be done to create models of individuals and their attempts before we can determine the impact of explicit strategies that focus on individual words.

As described in Equation (1), a second contributor to activation is the base-level activation of words in semantic memory. BLA is determined by the recency and frequency of access of an item. The following is the BLA equation that was used for our models:

$$BLA = \log\left[\sum_{i=1}^{n} t_i^{-d}\right] \tag{3}$$

where n is the number of activation boosts, t_i is the time since the i^{th} boost, and d is the decay factor. In this task, we expect that BLA plays a minor role if any because RAT problems are explicitly designed to avoid repetition of words and the associated priming effects provided by BLA. For these reasons, our initial model does not include BLA, but we explore it as a variant in the Model Variant Experiments and Evaluation section.

The final contributor to activation is noise. Soar and ACT-R differ in how they compute noise. The noise term in ACT-R is logistic, whereas in Soar it is Gaussian. A standard value for ACT-R's noise parameter is 0.4 (Taatgen, Lebiere, & Anderson, 2006), although it varies across models. While not exactly the same, we consider this value similar to setting Soar's noise magnitude to 0.73, which refers to the standard deviation of the Gaussian from which Soar draws its random noise term during activation computation. As with BLA, we introduce noise as a variant of our original model in the Model Variant Experiments and Evaluation section.

5 Experiments and Evaluation

In this section, we first describe the metric we use to evaluate alternative knowledge bases and models. We apply that metric to the knowledge bases independent of the process model. We briefly examine semantic distributional models and how they perform when directly solving the RAT. We then evaluate our model on all five knowledge bases. After determining that HBC and SWOWEN perform the best in terms of our evaluation, we use HBC and SWOWEN as the knowledge bases to explore model variants incorporating changes to association strength, base-level activation, and noise.

5.1 Evaluation Metric

To evaluate our model's performance in comparison to the human data from Bowden and Jung-Beeman (2003), we focus on question difficulty. Specifically, we ask whether questions that are difficult for humans to get correct are also difficult for the model to get correct. This was the most straightforward analysis given that the human data from Bowden and Jung-Beeman (2003) provides only aggregate human data for comparison. We evaluate the coefficient of determination (R^2) and the Mean-Squared Error (MSE) between the difficulty of RAT questions for humans and the difficulty of RAT questions for the models. As the measure of human difficulty, we use the percentage of subjects who got the correct answer. For a model's difficulty, we use whether the model produces the correct answer for a given problem. In order to easily compare these two metrics, we binned the 144 compound RAT problems into 12 groups of 12 RAT questions based on human correctness percentage for either 2, 7, or 15 seconds depending on the analysis used. For the 12 questions in each bin, we calculated the mean percentage of subjects who got the questions correct and then used that to determine the expected number of questions correct in each bin, as shown in Figure 2. The first

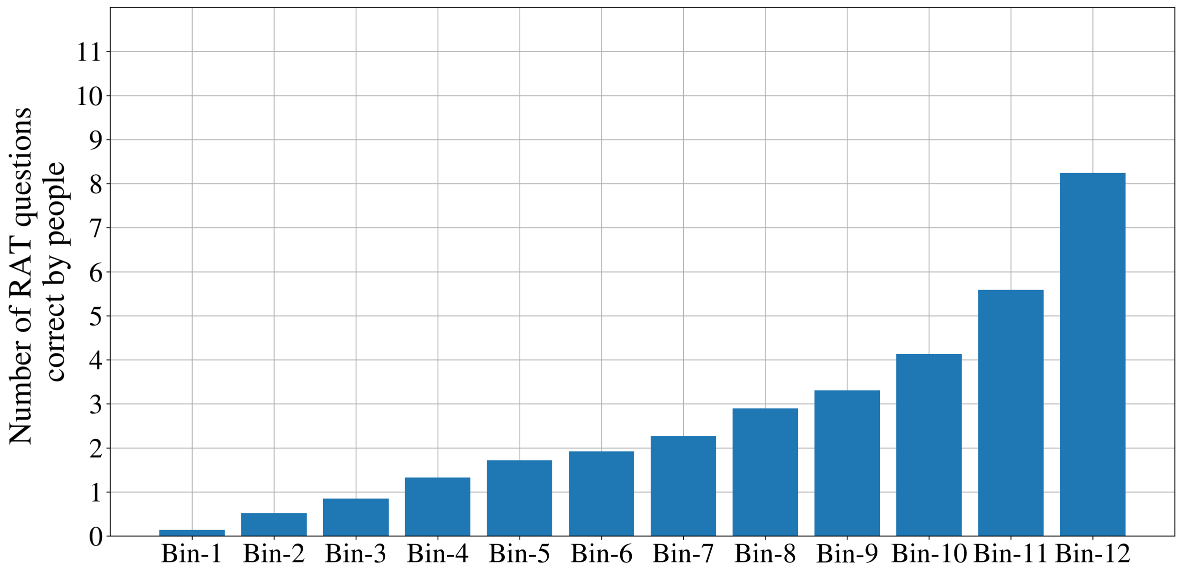

Fig. 2. The number of questions people are expected to get correct on average for each of the 12 bins, ordered from most difficult to least difficult.

bin has the 12 most challenging questions for humans (people are expected to get an average of 0.14 of those questions correct out of 12), while the last bin has the easiest questions (people are expected to get an average of 8.24 of those questions correct out of 12). In our evaluation, we compare the expected number of questions correct in a bin to the number of questions our models got correct for that bin for each of the 12 bins.

Using this approach, we attempt to provide a simple comparison between the human data and model data. We conceive of the human data as a sample mean that estimates the probability a human answers a given question correctly, i.e., "difficulty." But instead of comparing model performance on each question directly to such an estimate, we treat each bin of human data as representing the average of sample means from similar distributions. This permits linear fitting, which is easy to interpret and still sufficient for highlighting the major trends that appear in our investigation.

5.2 Baselines

5.2.1 Knowledge Base Baselines

Before evaluating the performance of our model, we evaluate the theoretical adequacy of the five knowledge bases for representing knowledge for modeling the RAT. This evaluation does not include any retrieval mechanisms, but examines whether solutions exist in a knowledge base and evaluates them on RAT problem difficulty as described above.

Table 4
Analysis of HBC for RAT in terms of directional links.

		7s		15s	
	# Correct	R^2	MSE	R^2	MSE
HBC Any Direction	109	0.68	41.83	0.57	32.10
HBC Correct Direction	82	0.44	19.52	0.30	14.76
HBC Forward Direction	55	0.84	4.51	0.74	2.78
HBC Backward Direction	58	0.26	8.34	0.13	9.23

In developing HBC, USF Norms, and SWOWEN the number of times that a word was presented was arbitrary, so it is possible that the number of associations from one word to its children is much greater than those of other words. Thus, we normalized the association weights between words based on the number of times the word was presented. Therefore, the association weights do not reflect frequency at which these word pairs occur in English.

Many of the knowledge bases, including HBC, have directional links between words. Most retrieval algorithms (including ours) are sensitive to those directions, so to evaluate the difficulty of a problem, we need to include separate analyses for the possible interactions between link directions and potential retrieval algorithms. We consider four possible methods for determining if a knowledge base contains the knowledge needed to correctly answer a RAT problem. Table 4, summarizes the results of our analysis for HBC.

The most inclusive method is to test whether an association exists in either direction between all prompt words and the solution word in each problem. With this criterion, as mentioned earlier, HBC has knowledge for 109 out of the 144 compound RAT problems, so that there are 35 problems for which the HBC is missing at least one association between a prompt word and the answer. A hypothetical model that answers all those problems correctly accounts for human difficulty on the RAT when humans are given 7 seconds with an R^2 of 0.68 and an MSE of 41.83. When humans are given 15 seconds, the results are an R^2 of 0.57 and an MSE of 32.10.

A stricter test of the knowledge base examines whether associations exist in the correct direction between every prompt word and the solution word in a problem. Different prompt-solution pairs may require an association in different directions. For example if the prompt word is “acid” and the solution word is “rain,” this method tests if there is an edge in the direction of “acid” to “rain” since “acid rain” is the association given by the solution. For the same RAT problem, this method also tests if the solution word “rain” has a edge in the direction of the prompt word “coat” since “rain coat” is the association given by the solution. With this criterion, HBC has correct associations for 82 out of the 144 compound RAT problems. A hypothetical model that gets exactly these 82 problems correct accounts for human difficulty on the RAT when humans are given 7 seconds with an R^2 of 0.44 and an MSE of 19.52. When humans are given 15 seconds, the results are an R^2 of 0.30 and an

Table 5
The best results of each knowledge base.

			7s		15s	
Knowledge Base	Best Direction	# Correct	R^2	MSE	R^2	MSE
COCA-TG	Both	94	0.20	31.25	0.19	22.84
Google Books	Forward	84	0.01	24.22	0.00	20.90
HBC	Forward	55	0.84	4.51	0.74	2.78
USF Norms	Correct	12	0.57	5.80	0.50	11.32
SWOWEN	Forward	26	0.65	2.11	0.49	5.85

MSE of 14.76.

A third method is to test if an association exists in the forward direction from every prompt word to the solution word in a problem. This is the direction spreading activation utilizes with associations. With this criterion, HBC has these associations for 55 out of the 144 RAT problems. A hypothetical model that answers all those problems correctly accounts for human difficulty on the RAT when humans are given 7 seconds with an R^2 of 0.84 and an MSE of 4.51. When humans are given 15 seconds, the result is an R^2 of 0.74 and an MSE of 2.78.

The last method is to test if an association exists in the backward direction from the solution word to every prompt word in a problem. With this criterion, HBC has these associations for 58 out of the 144 RAT problems. A hypothetical model that answers all those problems correctly accounts for human difficulty on the RAT when humans are given 7 seconds with an R^2 of 0.26 and an MSE of 8.34. When humans are given 15 seconds, the result is an R^2 of 0.13 and an MSE of 9.23. Although the number of correct answers is similar, the backward direction has a much lower R^2 because it is not getting the same problems correct as humans. It is more likely to get the difficult problems correct, and less likely to get the easy problems correct.

HBC forward has both the highest R^2 and lowest MSE for both 7 and 15 second conditions by large margins. These results suggest that the way HBC was collected and the existence of forward direction links is more predictive of the specific problems humans are able to answer than using the other links. This is also consistent with our model, which relies exclusively on forward links, modulated by association strength, to retrieve answers.

If we assume that the availability of links in HBC reflects a subset of human word association knowledge, any retrieval method that uses forward associations has solutions available to at least 55 RAT problems. With the retrieval algorithm, the model can get problems correct even if they are not fully in the knowledge base through the guessing mechanism. Even without the guessing mechanism, one question is why do humans get only 32.92 problems correct on average when given 7 seconds and 44.25 when given 15 seconds? What other factors hamper human performance? Our hypothesis is that other associated words with higher strengths interfere with the retrieval of the correct words.

In Table 5, we evaluate all knowledge bases (COCA-TG, Google Books, HBC, USF Norms, and SWOWEN) using the same metrics we used to analyze HBC. For each knowledge base, we include the results from their best direction with the corresponding number correct, R^2, and MSE for both 7 and 15 second human data. COCA-TG has bi-directional links so there is only way of analyzing if solutions to RAT problems exist, which we call "Both." With this method, COCA-TG has solutions to 94 RAT problems. When comparing to human data on the RAT, COCA-TG's highest R^2 is 0.20 and lowest MSE is 22.84. Google Books is a larger 2-gram word frequency knowledge base than COCA-TG and when using the forward direction it has solutions to 84 RAT problems. Google Books' highest R^2 is 0.01 and lowest MSE is 20.90. The poor correlations for COCA-TG and Google Books arise because they have almost even distributions of correct questions across the RAT difficulty problem bins, in contrast to the human results as shown in Figure 2. Compared to other knowledge bases, the main difference is that COCA-TG and Google Books have many solutions to RAT problems that people find difficult compared to only having many solutions to RAT problems that people find easy. In the Model Experiments and Evaluation section, we further explore if performance for COCA-TG and Google Books improves when used with our process model.

USF Norms' best direction was using the correct direction and with that method it has solutions for 12 RAT problems. Using those associations, USF Norms' R^2 is 0.57 and lowest MSE is 5.80. SWOWEN's best direction is forward and with that method it has solutions to 26 RAT problems. Using that direction, SWOWEN's R^2 compared to human data is 0.65 and its MSE is 2.11.

HBC is larger and more comprehensive than both SWOWEN and USF Norms and this analysis of the knowledge bases suggests that HBC has the best potential for modeling humans on the RAT. However, this analysis does not include a retrieval algorithm nor the rest of the model. In the Model Experiments and Evaluation Section, we explore how each of these knowledge bases perform with Soar retrieval mechanisms and the algorithm for solving the RAT.

5.2.2 Language Model Baselines

In addition to the cognitive process model we created, we also investigated distributional semantic models (DSMs) for the RAT. Specifically we examined how word2vec and GloVe performed on the RAT with a n-nearest neighbor algorithm. Word2vec "provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words" (Mikolov, Chen, Corrado, & Dean, 2013). For this work we used a model that was pre-trained on google news. The model consists of 3 million 300-dimensional vectors. In contrast, "GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space" (Pennington, Socher, & Manning, 2014). For this work we used a model that was pre-trained on Wikipedia 2014 and Gigaword 5. The model consists of 400K 300-dimensional vectors.

The stand alone models for both word2vec and GloVe do not use Soar to solve the RAT problems. Instead, for each RAT item, the model sums or averages together the vectors of the three prompt words. The model then finds the closest X words, using euclidean distance, to that summed or averaged vector. For X=1, the number of correct answers to the 144 RAT problem using word2vec and GloVe is 0. That is, neither models finds any correct answers to the RAT problems for its best answers. Beyond the best answers, a model can look at more than just the closest word by exploring X > 1. However, once there are multiple options, the model can only differentiate them by choosing the word that is closest to the original sum or average, which still leads the models to getting no questions correct (same as X = 1). We implemented an additional mechanism that searches through the X options for the sought after solution to each specific RAT item. If the search is successful, we count that as getting the RAT question correct. Note that this requires knowledge of the solutions, which is not available to the models. This knowledge is included to evaluate whether the solutions are potentially available. The Soar models do not require a search mechanism such as this because they already assess if a possible solution is connected with the three prompt words. Such a mechanism is not possible for the word2vec and GloVe models.

We explored a series of X's as well as summing versus averaging the vectors. For word2vec, the best results get 6 RAT problems correct when expanding X out to 50, indicating that word2vec does not contain the information required for modeling the RAT. For GloVe, when compared to 7 second human data, the best model sums the vectors and checks the closest 8 words. This combination has an R^2 of 0.83 and a MSE of 1.19 with a total of 26 problems correct. When compared to 15 second human data, the best model sums the vectors and checks the closest 10 words. This combination has an R^2 of 0.64 and a MSE of 4.23 with a total of 29 problems correct. As we shall see in the next section, the results for these incomplete stand alone models do not compare favorably with those of the Soar models, and will not be investigated further.

5.3 Model Experiments and Evaluation

In this section, we evaluate our Soar process model using the five knowledge bases. An important aspect of our model is that it makes multiple attempts until it retrieves a result that has associations with all prompt words or hits a limit on the number of attempts. The parameter that determines the number of attempts needs to be determined for the two different amounts of time the humans were given (either 7 or 15 seconds). As stated earlier, we do not model all the detailed processing required in this task (such as reading prompt words and writing answers), so we avoid making a priori predictions as to the number of attempts a subject can make in 7 or 15 seconds, although we predict that more attempts are possible in 15 seconds than 7 seconds. Instead, we take an empirical approach to determine the number of attempts that best model the human data.

For all knowledge bases, we ran experiments to find the best attempt number given 7 and 15 seconds. As an example of what we found, Figure 3 shows the results of the model with the HBC

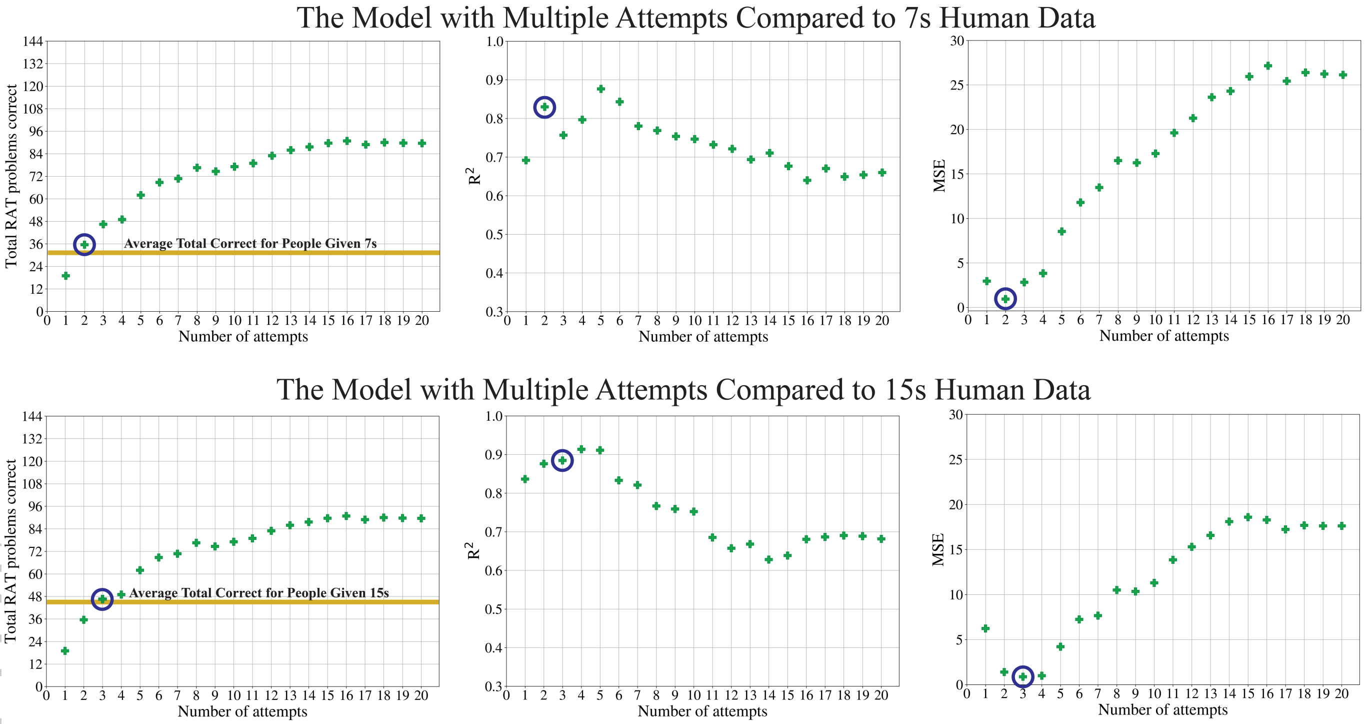

Fig. 3. The model with HBC using 1 through 20 attempts on the RAT presented in terms of: total number of problems correct, R^2, and MSE for 7 second (top) and 15 second (bottom) human data.

knowledge base using spreading activation with association strengths (without BLA or noise). The number of attempts is varied from 1 to 20. The left-most graphs show the total number of problems that the model gets correct, which increases as more attempts are made (as expected). The middle graphs shows R^2 (higher is better), which varies between 0.63 and 0.91. The graphs on the right shows MSE (lower is better), which initially dips and then increases as the model gets more questions correct than humans do.

We use MSE to choose the best attempt number. Thus, for 7 seconds the best attempt is 2 attempts (circled on the graph for each metric). For 15 seconds the best attempt for number of problems, is 3 attempts (circled on the graph for each metric), with 4 attempts being a close second. Our goal is for the model to match human behavior and not get the most RAT questions correct. Therefore the best attempt achieves a low MSE, indicating that the questions people struggled with are the same our model had difficulty with.

If the model does not find a possible solution, it makes a guess, selecting from words that are connected to two prompt words. These guesses are sometimes correct, even if the model does not recognize that. Additionally, the model can believe it found the correct answer, i.e., it found a word with links to all three prompt words, but in reality that word is not the sought after response. This is true of knowledge bases, such as Google Books, that have large number of associations that are not the kind used in the RAT. Such a result is reported as a false positive. Note if the model believes

Table 6

Results of the five knowledge bases with spreading and association strengths for the best number of attempts.

	7 Seconds				15 Seconds			
	Attempts	# Correct	R^2	MSE	Attempts	# Correct	R^2	MSE
COCA-TG	1	16.0	0.78	3.15	3	41.0	0.68	2.90
Google Books	2	13.0	0.65	4.69	4	17.0	0.65	9.22
HBC	2	35.5	0.83	0.93	3	46.4	0.88	0.88
USF Norms	5	27.7	0.86	0.92	5	27.7	0.78	3.42
SWOWEN	5	34.5	0.97	0.29	8	48.6	0.91	1.33

Table 7

Results for the five knowledge bases in terms of number of guesses and false positives for the best number of attempts (see Table 6 for best attempt).

	7s			15s		
	# Correct	Guesses	False Positives	# Correct	Guesses	False Positives
COCA-TG	16.0	6.0	1	41.0	18.0	7
Google Books	13.0	3.0	69	17.0	3.0	79
HBC	35.5	2.5	4	46.4	3.4	5
USF Norms	27.7	18.7	1	27.7	18.7	1
SWOWEN	34.5	14.5	2	48.6	19.6	4

it got the correct answer, it will not make additional attempts. Guesses and false positives have not been previously explored in RAT models. Unfortunately, the information needed to assess whether a person reported a guess or false positive is not included in the RAT data we use for these experiments, so it is not possible to determine if a correct answer resulted from a confirmed answer or a guess. Therefore, they are interesting metrics of the model to consider but currently they do not provide evidence of more or less human like behavior.

Table 6 shows the results of the five knowledge bases with spreading activation and association strengths using the best number of attempts for that knowledge base. Table 7 shows the number of RAT questions correct, the number of guesses, and the number of false positives for the best attempt for each knowledge base. The best model uses SWOWEN as its knowledge base with an R^2 of 0.97 and MSE of 0.29 compared to 7 seconds human data and an R^2 of 0.91 and MSE of 1.33 compared to 15 seconds human data. However, the model using SWOWEN has a high number of guesses with 14.5 for 7 seconds and 19.6 for 15 seconds respectively. In other words, of the 34.5 RAT problems the model got correct given 5 attempts, 14.5 of them are guessed. This is in strong contrast to the model using HBC, which only guesses 2.5 times for 7 seconds and 3.4 times for 15 seconds. The model with HBC has an R^2 of 0.83 and MSE of 0.93 compared to 7 second human data and an R^2 of 0.88 and MSE of 0.88 compared to 15 second human data. USF Norms does well compared to 7 seconds human data with an R^2 of 0.86 and MSE of 0.92 and has 18.7 guesses. However, compared to the 15 seconds human data, this model has an R^2 of 0.78 and MSE of 3.42 and has 18.7 guesses.

Neither of the knowledge bases built from word co-occurrences, Google Books and COCA-TG, do as well as HBC, SWOWEN, or USF Norms in terms of R^2 and MSE. Google Books gets only 13 questions correct given 2 attempts but it has 69 false positives (and 2 guesses). In other words, this model thinks it is getting 79 questions correct (total - guesses + false positives) while it is actually getting only 13 correct. COCA-TG does better, but its performance is far below the other knowledge bases. The problem is that these knowledge bases have large numbers of links between words that are *not* related to compound words or phrase associations and those links lead to incorrect answers. For example, given the RAT item "show", "life", and "row," the model using Google Books finds "cycle" on its first attempt. Within Google Books, this word has associations to all the prompt words, so it reports it as an answer, whereas the sought after response is "boat."

The final observation is that the best models for 7 seconds data use fewer attempts than the best models for 15 seconds data. This is true for all model knowledge base combinations except for USF Norms, whose best results are for 5 attempts for both 7 and 15 seconds data.

To summarize, we examined two different types of knowledge bases, word frequencies (COCA-TG and Google Books), and recorded word associations (HBC, USF Norms, and SWOWEN). The best models used recorded word associations. Specifically, the best were HBC and SWOWEN when matching both 7 and 15 second human data. SWOWEN had the best R^2 and MSE but also used many guesses, which makes up for the fact that it is missing some links between prompt and solution words. HBC did almost as well as SWOWEN and used few guesses. A direct visual comparison of the knowledge bases in terms of R^2 and MSE can be seen in Figure 4. In the next section, we explore variations of these top two models to better understand what aspects are important for matching human behavior.

5.4 Model Variant Experiments and Evaluation

In this section, we investigate variations of the models relative to the different aspects of the activation formula as originally shown in Equation (1), including association strength, BLA, and noise.

To test the influence of association strength, we created model variants where all associations strengths are equal. To test the influence of BLA, we created models where items in semantic memory have an initial activation derived from word frequency. To test the influence of noise, we created models with varying levels of noise in computing activation. Our goal is to further understand what aspects of activation influence the performance of our models on the RAT.

While we explored variants with all the knowledge bases, we only present variants with our top two performing knowledge bases, HBC and SWOWEN. In no cases did variants using the other knowledge

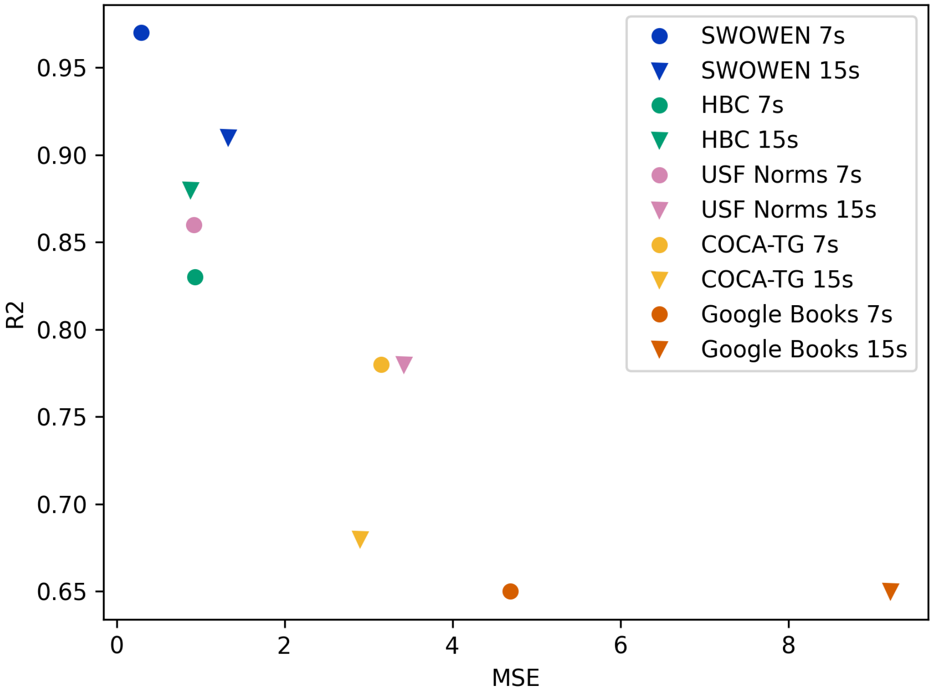

Fig. 4. Best model performance for each knowledge base, plotted according to R^2 and MSE.

bases perform better than (or even close to) the performance achieved with HBC or SWOWEN.

5.4.1 Activation Spread without Association Strength

To determine the impact of association strength on retrieval, we developed a variant where all association strengths are the same. Effectively, this reduces the spreading activation computation to performing a count for each recipient for how many sources of spread have associations to that recipient. This is similar to the approach Kajić et al. (2017) used with their neuron spiking model in that they did not incorporate association strength when retrieving potential solutions.

As shown in Table 8, removing the effects of association strength influences the number of questions answered correctly, R^2, and MSE for both HBC and SWOWEN. The table shows the results without association strength using the same number of attempts used earlier, as well as the number of attempts that were best given no variation in association strength. R^2 improves in only one case (HBC 7 seconds) when association strength is removed. However, removing association strength increases the number of items correct and increases MSE.

This suggests that the strength of the associations contributes to the difficulty of the RAT problems because for many of the problems, the correct answer is not the one most highly associated with the prompt words. Instead, other (incorrect) words have high enough association strengths from one or two of the prompt words so that they are retrieved first, even if they are not associated with all the

Table 8

Each knowledge base is presented with the base model given its best attempt, with no association strength given the base model's best attempt, and with no association strength with that best attempt (indicated with an *).

	7 Seconds				15 Seconds			
	Attempts	# Correct	R^2	MSE	Attempts	# Correct	R^2	MSE
HBC	2	35.5	0.83	0.93	3	46.4	0.88	0.88
HBC No Assoc	2	54.9	0.88	5.34	3	68.9	0.78	6.63
HBC No Assoc*	1	47.0	0.91	3.37	1	47.0	0.86	2.77
SWOWEN	5	34.5	0.97	0.29	8	48.6	0.91	1.33
SWOWEN No Assoc	5	52.0	0.66	4.77	8	54.9	0.68	3.38
SWOWEN No Assoc*	2	49.8	0.71	3.66	2	49.8	0.77	2.14

prompt words. In contrast, the correct answers have low associations with all prompt words, as was the case in the example in Table 3, and require additional attempts to be retrieved. We conclude that association strength is an important component in modeling human performance.

5.4.2 Base-Level Activation

In the models described so far, the initial activation of all words in long-term memory are the same. They do not include any BLA, which is computed from the frequency and recency of access from long-term declarative memory as originally described in Equation (3). The RAT is explicitly designed to avoid short-term base-level effects by not repeating any of the solution words. So when we created a model that includes short term base-level effects, we did not expect see significant changes to the results. However, there is the potential of long-term effects where common words (with high frequency) are more likely to be retrieved and thus can affect the difficulty of problems.

Table 9

Each knowledge base is presented with the base model and with BLA. For both HBC and SWOWEN, the number of best attempts (2 and 5) did not change when BLA was included.

	7 Seconds				15 Seconds			
	Attempts	# Correct	R^2	MSE	Attempts	# Correct	R^2	MSE
HBC	2	35.5	0.83	0.93	3	46.4	0.88	0.88
HBC w/BLA	2	31.5	0.89	0.59	3	41.1	0.87	0.99
SWOWEN	5	34.5	0.97	0.29	8	48.6	0.91	1.33
SWOWEN w/BLA	5	34.1	0.89	0.76	8	47.4	0.87	1.56

In an attempt to determine whether BLA influences the results, we initialized the BLA values for words in semantic memory using word frequency data from COCA (Davies, 2008). Note that BLA depends on a pattern or history of usage, and not just frequency. However, rather than attempting to implement a full model of word usage history, we instead created an approximation. All words were initialized with BLA as if they had a number of accesses equal to their word frequency data. These accesses were modeled by setting the time of access for all elements to the same time-step in the past for the agent. This is not a psychologically realistic implementation but is functional because it allows BLA from before the task to depend on word usage without having large short-term transient decays. A free parameter was the amount of time into the past that this initial access data was set to. Variations in the value of this parameter had little impact on results. We used word frequency initialized 1000 time steps in the past for our models. In most cases, responses did not change because the boost from receiving spread from prompt words dominated differences in base-level from word frequency, and any changes did not have a significant impact on the performance measures.

The results for our top two knowledge bases, HBC and SWOWEN, are presented in Table 9. Adding BLA had minor changes to the results for both HBC and SWOWEN. We also added BLA to the model that uses Google Books (not shown) to ensure that it did not significantly improve its performance. In fact, it went from 13 RAT problems correct without BLA to 1 correct, allowing up to 10 attempts. Its number of false positives also went up from 69 with the base model using 2 attempts to 115 with BLA model using 2 attempts. We conclude that BLA is not an important component for human difficulty in this version of the RAT.

To further investigate this claim, we examined how well word frequency alone predicts solutions to the RAT problems without a model. For HBC and SWOWEN, for each prompt word, we examined the outgoing links and found the rank of the solution word in terms of word frequency in relation to all other linked words. We then ordered all words by their rank and binned them into 12 bins. We then averaged the ranks across each of the 12 bins as a potential model of problem difficulty. We then compared those ranks to the human data for each bin. For SWOWEN, given 7 and 15 seconds, it had an R^2 of 0.03 for both. For HBC, given 7 and 15 seconds, it had an R^2 of 0.09 and 0.10. This is further evidence that word frequency does not play major role in determining difficulty of RAT problems.

5.4.3 Activation Calculation with Noise

The final variant involves adding noise to the calculation of activation in the initial model. We examined four noise magnitudes: 0.0, 0.5, 1.0, and 1.5. One major observation is that as noise is added, the total number of RAT problems answered correctly decreases linearly. For example given 2 attempts with the HBC knowledge base, as shown in Table 10, the total number of correct responses decreases from 35.5 with no noise to 31.3, 24.9, and 19.0 as we increment by 0.5 noise. This has a strong negative correlation of -0.99. The same trend can be seen with SWOWEN in Table 11.

Table 10

Total RAT problems correct given varying levels of noise with HBC.

	1 attempt	2 attempts	3 attempts	4 attempts	5 attempts
0.0 Noise	19.0	35.5	46.4	49.0	63.03
0.5 Noise	17.5	31.3	42.4	51.0	57.8
1.0 Noise	13.7	24.9	34.6	42.4	49.0
1.5 Noise	10.0	19.0	26.6	33.0	38.4

Table 11

Total RAT problems correct given varying levels of noise with SWOWEN.

	1 attempt	2 attempts	3 attempts	4 attempts	5 attempts
0.0 Noise	11.0	16.9	23.7	30.2	34.5
0.5 Noise	9.5	16.5	21.5	31.5	34.0
1.0 Noise	8.1	14.3	19.9	25.3	29.3
1.5 Noise	6.3	11.1	15.2	20.0	23.8

Thus, as the noise increases, so does the likelihood of retrieving other words that are not solutions. This is especially true for RAT solutions that have weak connections with the cue words, which in general are the more difficult problems. This means that the model is getting fewer difficult questions correct which in turn increases the R^2 and decreases the MSE. This can be seen with the model using HBC and noise 0.0 to noise 0.5 with 2 attempts see Table 12. The R^2 goes from 0.83 to 0.94 and the MSE goes from 0.93 to 0.37. However, as even more noise is added, the model starts to also get more easy questions wrong as well and the R^2 drops as the MSE rises. Not shown in a table, but as noise increases from 1.0 to 1.5 with 2 attempts, R^2 goes from 0.94 to 0.88 and the MSE goes from 1.11 to 2.93.

For SWOWEN, getting fewer RAT solutions that have weak connections with the cue words does not mean only getting fewer difficult questions correct. It also means getting fewer easy questions correct. When looking at the SWOWEN model with noise 0.0 to noise 1.0 with 5 attempts, the R^2 barely changes from 0.97 to 0.95 and the MSE also barely changes from 0.29 to 0.33. This is even though the total RAT questions correct drops from 34.5 to 29.3 which means the change in each bin is small. This can be seen in Table 13. However, as even more noise gets added, the R^2 stays similar but the MSE increases. Not shown in a table but for the SWOWEN model given 5 attempts for noise 1.0 to 1.5, the R^2 goes from 0.95 to 0.96 and the MSE goes from 0.33 to 1.07. This shows that as noise increases the model is getting fewer easy and difficult questions correct keeping the same relation to human difficulty in terms of R^2 but eventually increasing the MSE as the total questions correct for each bin decreases below what people get correct.

One way to combat the decreasing number of correct answers that occurs because of noise is to increase the number of attempts. This gives the model more chances of retrieving the solution and empirically increases the number of correct answers for both easy and hard problems. Going from 1 attempt to 5 attempts with no noise, the model using HBC increases from 19.0 questions correct to 62.7 questions correct, forming a positive linear relationship with a correlation of 0.98. Achieving the best model performance requires balancing noise with number of attempts, both of which affect the R^2 and MSE.

As shown in Table 12, adding 0.5 noise to the model using HBC does lead to improved performance for both the 7 second and 15 second data. The differences in performance for 0.5, 1.0, and 1.5 noise are minimal.

The impact for SWOWEN (Table 13) is even less clear with 0.0 noise being the best for 7 seconds, but only marginally, whereas 0.5 is best for 15 seconds. The performance of the models in most cases is so good (R^2 above 0.93 and an MSE below .50) that attempting to find the absolute best model in terms of combinations of noise and attempts is a bit of an academic exercise. The most important lesson is not that noise improves or does not improve the ability of the model to match human data, but that by increasing the number of attempts, the model can compensate for noise if it exists.

Table 12
Result summary from the initial model with noise using HBC.

	7s				15s			
	Attempts	# Correct	R^2	MSE	Attempts	# Correct	R^2	MSE
Human	–	32.9	–	–	–	44.3	–	–
0.0 Noise	2	35.5	0.83	0.93	3	46.4	0.88	0.88
0.5 Noise	2	31.3	0.94	0.37	3	42.4	0.94	0.49
1.0 Noise	3	34.6	0.92	0.44	4	42.4	0.94	0.43
1.5 Noise	4	33.0	0.93	0.49	6	44.1	0.93	0.50

Table 13
Result summary from the initial model with noise using SWOWEN.

	7s				15s			
	Attempts	# Correct	R^2	MSE	Attempts	# Correct	R^2	MSE
Human	–	32.9	–	–	–	44.3	–	–
0.0 Noise	5	34.5	0.97	0.29	8	48.6	0.91	1.32
0.5 Noise	4	31.5	0.90	0.63	6	42.0	0.98	0.29
1.0 Noise	5	29.3	0.96	0.33	9	42.0	0.95	0.47
1.5 Noise	8	32.1	0.94	0.32	12	40.2	0.95	0.44

As shown in Tables 14 and 15, we also created a set of models that incorporated combinations of all the variants for HBC and SWOWEN to determine if there are any important interacts among

Table 14
Result summary from BLA model and no association strength model and 0.5 noise using HBC. All models are presented with their best attempt.

	7s				15s			
	Attempts	#Correct	R^2	MSE	Attempts	# Correct	R^2	MSE
BLA Noise 0.5	2	30.6	0.90	0.61	3	42.0	0.93	0.52
BLA	2	31.5	0.89	0.59	3	41.1	0.87	0.99
No Assoc Noise 0.5	4	40.1	0.90	1.07	5	46.5	0.87	1.05
No Assoc	1	47.0	0.91	3.37	1	47.0	0.86	2.77

Table 15
Result summary from base-level initialized model and no association strength model and 0.5 noise using SWOWEN. All models are presented with their best attempt.

	7s				15s			
	Attempts	#Correct	R^2	MSE	Attempts	# Correct	R^2	MSE
BLA Noise 0.5	5	33.1	0.95	0.33	8	44.6	0.93	0.80
BLA	5	34.1	0.89	0.76	8	47.4	0.87	1.56
No Assoc Noise 0.5	3	29.6	0.72	1.64	6	42.1	0.72	1.98
No Assoc	2	49.8	0.71	3.66	2	49.8	0.77	2.14

model variants. The models vary by including 0.0 or 0.5 noise, BLA or not, and association strengths or not. All results are shown with the best number of attempts for that specific model. The results here are consistent with those reported above. Adding 0.5 noise to these models mostly improved their performance compared to the versions without noise, usually decreasing MSE while having little to no impact on R^2, with the number of best attempts either staying the same or increasing. Also, none of these combinations end up dominating the base models described above where noise and association strength are included but BLA is not.

5.5 Statistical Significance

We found statistical significance regarding the choice of knowledge base, the use of association strengths, and the relationship between noise magnitude and best number of attempts. But given the high R^2 and low MSE values among the best models, it is not surprisingly that some of the differences between the model variants are not statistically significant, especially for models that include noise.

The models using HBC or SWOWEN have lower MSE than other models with a probability that this result occurs by chance (p-value) of $p < 0.01$ for both 7 seconds and 15 seconds data. Among the models that do not use either HBC or SWOWEN, it was difficult to find any significant trends.

Among the model variants using either HBC or SWOWEN, we find three significant results. First, models with association strength are better than models without ($p < 0.01$). The second result is more nuanced. Adding increasing amounts of noise reduces the number of problems a model gets correct. Also, increasing the number of attempts increases the number of problems a model gets correct. We hypothesized that doing both could potentially change how well a model corresponds to human performance while achieving the same total number correct. Instead, we often found no statistically significant difference between a model with a small amount of noise and a model with more noise and compensatory additional attempts. These results hold across a range of model parameter choices, but the third significant result was sensitive to parameter choices. When using the best model parameters for HBC and for SWOWEN, the best model using SWOWEN is significantly ($p < 0.01$) better than the best model using HBC. Appendix B describes our method for determining statistical significance.

6 Related Work

There has been extensive research on the RAT (Wu et al., 2020); however it can be difficult to directly compare our results to much of the prior research because of differences in data sets, metrics for comparing human data, and modeling approaches. Below we review the most relevant previous research on how people solve the RAT, with emphasis on the unique aspects of our work.

In our work, we focus on matching aggregate human performance to computational models in terms of problem difficulty. Along similar lines, but through the analysis of individual differences, Marko, Michalko, and Riecansky (2019) found that difficulty is related to the remoteness of the associations. Other work in individual differences on the RAT study the potential connections between individual performance on the RAT and other cognitive skills or characteristics, including convergent versus divergence thinking tests of creativity (Lee Bae, Huggins-Manley, & Therriault, 2014), and speed of retrieval and creativity (Benedek & Neubauer, 2013).

Kajić and Wennekers (2015) created a neural network model of the RAT that solves the problems using spreading activation and a winner-take-all mechanism. Their model, like ours, examined using multiple attempts to solve RAT problems. They showed that for their method using 6 attempts led to a similar total correct as people given 15 seconds to solve a subset of the 144 RAT problems by Bowden and Jung-Beeman, 28.8% versus 28.2% correct. Their work split the RAT problems into three bins based on difficulty and showed that more attempts are needed for questions in the harder bin. They also showed that association strengths affects problem difficulty for the model by using a threshold for connections in the knowledge base and showing a change in the number of problems correct for each difficulty bin. Our work differs by providing a finer-grain analysis of factors influencing retrieval, such as BLA and noise, and a finer-grain analysis of problem difficulty (12 bins vs. 3). Our work also examines five different potential knowledge bases while theirs only considered one, USF Norms. We also differ in that we use a different modeling methodology and utilize existing memory retrieval mechanisms available with cognitive architectures.

Kajić et al. (2017) created a spiking neuron model of word associations to solve the RAT using iterative retrievals. Their work showed how you can represent task knowledge with biologically realistic spiking neurons. Based on previous research, they found that USF Norms provided the best word association data for modeling the RAT when compared to Google Books (Kajić et al., 2016; Lin et al., 2012; Nelson et al., 1998). This led them to create a model that uses USF Norms when retrieving potential RAT solutions. Their work used a set of 25 RAT problems previously studied by Smith et al. (2013). They compared their retrievals to human data on the 25 RAT problems where people were given 2 minutes to solve each problem, whereas in the experiments we model, people have either 7 or 15 seconds. They modeled individual participants by varying the knowledge base used in runs of their model. They found that average problem accuracy correlates with human data with an r = 0.49. However they also found that for 14 out the 25 RAT problems, their model differed significantly from the human responses. For those problems the model had either a significantly easier or harder time answering than people. Our results show a close relation for difficulty between our model and the human results. They also did not evaluate the impact of association strength to influence retrieval, which we found is an important component, nor did they explore the impact of different levels of noise. Similar to our model, their model filtered retrieved words to give only viable potential solutions to each RAT problem. Their best model used USF Norms for retrieving potential solutions and Google Books for filtering. In contrast to using two different knowledge bases, our model uses a single knowledge base for both retrieval and filtering.

Moss (2006) used RAT problems to analyze how "open goals" influence performance. In their work, a previously seen but unsolved RAT problem was considered an open goal (Moss, 2006). They performed multiple human studies as well as created a model for solving the RAT in the cognitive architecture ACT-R. Unlike our approach, they focused on how the difficulty of problems varied after a subject had been given a hint or had previously failed to solve the problem (leading to an open goal) as opposed to attempting to model problem difficulty across RAT problems. They found that giving a hint for RAT problems helped people with both unseen and unsolved problems. Their work also differs from ours in that they used a hand-crafted database of words and their connections for their model.

Olteţeanu and Falomir (2015) created a model of human performance, called comRAT-C, within the CreaCogs architecture. It was inspired by their account of creative problem solving which posits two extremes of behavior: "creative search" and "productive representation construction processes." The "creative search" extreme is implemented in comRAT-C using associational links to search a knowledge base for a solution to a RAT problem. Their knowledge base, called RAT-KB, uses the most frequent 2-grams from COCA (Davies, 2008). In a separate analysis (Olteţeanu & Schultheis, 2017), they state that the difficulty of the RAT depends on "(i) the frequency of a query-answer association, as a form of associative strength and (ii) the ratio between such an associative strength and the number of answer associations." We interpret these factors as being analogous to how association strength (Anderson & Pirolli, 1984) and fan (Anderson, 1974) are theorized to govern retrieval difficulty. Their comRAT-C model calculates the probability that a word is the answer, given a prompt word: $p(w_{ans}|w_q) = \frac{fr(w_q \rightarrow w_{ans})}{\sum_{k=1}^{n} fr(w_q \rightarrow w_k)}$. They then calculate the probability that a word is the answer with the expression: $p(w_{ans}) = \frac{1}{3} \cdot [p(w_{ans}|w_a) + p(w_{ans}|w_b) + p(w_{ans}|w_c)]$.[3] They measure how predictive $p(w_{ans})$ is for the percentage of human participants that solve the task given 7, 15, and 30 seconds. They report finding a correlation R value of 0.45, 0.41, and 0.49 respectively. Instead of comparing the raw frequency of word association pairs to human data for each problem, as described above, we compare question difficulty for our computational process models against question difficulty for humans. We contend that this is a better measure of performance of each model as it takes into account competition of retrieval in addition to frequency of word association pairs. We expand on their model in three ways: by using a more extensive knowledge base, by creating a computational process model implemented in a cognitive architecture, and by expanding the evaluation of the results on model variants, including the influence of multiple retrievals and noise.

Other work has taken a network science approach to solving the RAT (Valba et al., 2021) that does not include a running cognitive process model. In this work, they calculated the probabilities of selecting words and then ran a simulation a number of times. Similarly to how we and previous work use spreading activation, they use a random walk with attraction to stimuli algorithm. They found that their results match human difficulty with an R of 0.74. However, they only used SWOWEN as a knowledge base while our work looks at five different knowledge bases. Their work found that the hardness of RAT problems relates to where the solution is in the semantic network compared to the prompt words and the strengths of the connections.

Davelaar (2015) explored what makes RAT problems difficult and what leads to good strategies for solving RAT problems from a data science perspective. They found that the best strategies examine all three prompt words and their surrounding words. This is similar to how our model uses spreading from the prompt words to produce potential solutions to the RAT. However, they did not create a computational model and compare its predictions to human data.

Our efforts build on these previous works to model the RAT. Across all the related work, our work differs in four major ways. First we look at five different knowledge bases and evaluate them as potential fits for modeling the RAT. Second, we use default cognitive architecture mechanisms to create our model. Third, we investigate and systematically explore how variants, such as noise, number of attempts, association strength, and base-level activation influence behavior. Fourth, we compare relative difficulty of RAT problems based on human data to our modeling results, achieving very good fits to human performance.

[3] The formula reported in their paper: $p(w_{ans}) = \frac{1}{3} \cdot p(w_{ans}|w_a) + p(w_{ans}|w_b) + p(w_{ans}|w_c)$, appears to be in error.

7 Discussion and Analysis

A major challenge in modeling human performance on the RAT is not just correctly answering many problems, nor even correctly answering the same number of problems that humans correctly answer, but correctly answering the *same problems* that humans correctly answer, while also not correctly answering the problems humans have difficulty with. We developed a base model that has three components: a knowledge base, a retrieval algorithm, and a processing model for iteratively retrieving and testing possible answers. We explored variants of all three components and found that changes to each of these components influenced the difficulty of problems and the model's ability to match human behavior. Those results were consistent across the 7 and 15 second data sets. When the model is applied to the 2 second data, using 1 attempt, the SWOWEN database, and noise of 1.0, it gets 8.12 correct (compared to 8.0 for humans). It has R^2 of 0.88, and MSE of 0.26, making it comparable to the performance on the other data (better MSE but worse R^2).

In the remainder of this section, we review each of those three components in turn, summarizing the lessons learned from our explorations of the baselines and their variants.

7.1 Knowledge Bases

We explored two methods for creating a knowledge base, resulting in five knowledge bases. The first method used word co-occurrences in English text, which was used to create Google Books and COCA-TG. Neither of these performed well in comparison to the best knowledge bases HBC and SWOWEN. Our explanation is that the co-occurrence associations do not capture the type of associations humans use in retrieving possible solutions. As a result, they have difficulty retrieving the correct answers. Instead, they often retrieve words that have associations with all three prompt words, but not the type of compound word associations tested in the RAT. Google Books is by far the largest knowledge base we considered, and our results show that just having more words and more associations does not lead to a better model.

The second and best method for creating a knowledge base involved gathering word associations directly from people using free form generation of directional associations from specific words. The three knowledge bases (HBC, USF Norms, and SWOWEN) created using this method differed in the number of words and associations, and the exact values for R^2 and MSE when compared to human performance, with SWOWEN edging out HBC. Association strength was critical to matching human performance. Without it, these model got many more problems correct than humans, suggesting that association strength makes the problems harder by making it difficult to retrieve the correct answers. That is, the RAT problems often require retrieval of lower-valued associations.

The performance of the models based on this final class of knowledge bases dominated those developed using the alternative methods suggesting that when solving RAT problems, people draw on the same associations as they do when asked to freely recall associations (HBC, USF Norms, and

SWOWEN) and do not draw on associations purely from word frequency (COCA-TG, Google Books). The simple conclusion is that people use the same associations in generating possible answers to the RAT as they do when asked to generate associations using free recall. Not only are the associations the same, but our results suggest that the distributions of association strengths are similar so that taking the frequency of associations of a word to other words is comparable to the probability of retrieving an associated word.

We also created stand alone models from the distributional language model baselines to determine if they include the appropriate associations required to perform the RAT. These are not complete models as they generate only potential solutions, and have no means to determine which of those is the correct answer. If we assume there is an additional mechanism for choosing the correct answer from the potential results, the results from using word2vec are poor, whereas using GloVe produces moderate results.

Word association knowledge bases have been compared to distributional semantic models before (Kumar, Steyvers, & Balota, 2021). Similar results are found that for many tasks word association knowledge bases perform better than distributional semantic models. Kumar et al. identified two possible reasons for this difference. First, that word associations gathered from people capture different modalities than just looking at text. For example, visually, the word "yellow" is highly related to the word "banana." However, a text based model may instead find "apple" or "orange" highly related to "banana." Second, collecting word associations is a semantic retrieval task on its own which makes it suitable for modeling other semantic retrieval tasks.

From our own work we conclude three things are necessary for having a good knowledge base for modeling the RAT. One, it needs the content and connections of the RAT problems. For example, USF Norms has the fewest number of RAT problems of all the gathered word association knowledge bases and out of the three gathered word association knowledge bases it performs the worst. Two, the knowledge base needs other content that competes and interferes with the correct content. When removing association strength, the models got more RAT problems correct but their ability to match human data decreased. Three, the association strengths and variation in connections needs to correspond to what we see in human difficulty. For example, Google Books has many RAT problems in its knowledge base but fails to get as many problems correct as people because the words that have strong connections to potential solutions do not correspond to what humans recall.

Although the models using word association knowledge bases perform surprisingly well on the RAT, these knowledge bases contain only a small subset of the knowledge that we assume is encoded in human long-term memories. It is more likely that humans have much more extensive knowledge bases, closer in magnitude to Google Books (which did have the necessary associations for all RAT problems, but was unable to effectively retrieve them). This raises the question as to how humans restrict themselves to searching the associations that are relevant to the RAT and ignore the additional irrelevant ones. Does activation spread only through these types of link or do humans have additional mechanisms to focus spread to only the relevant associations? These are open question beyond the scope of this paper and worthy of future research.

7.2 Retrieval Algorithm

The basic model relies on the Soar implementation of ACT-R's declarative memory retrieval mechanism, which uses a form of spreading activation from prompt words along forward links to determine possible responses. This type of retrieval does not guarantee that a retrieved word has associations with all prompt words. Instead, it returns an answer with the highest total activation, which is often a word that is strongly associated with a single prompt word. This research provides evidence that the ACT-R retrieval algorithm is consistent with the mechanisms used by humans, but it does raise the issue identified above as to how spreading is limited to the associations relevant to the RAT.

In additional to spread, we investigated the two other components of the retrieval algorithm, namely base-level activation and noise. BLA in our experiments had minimal impact, which was expected given that the problems were designed to avoid repetition. We investigated noise by creating variants with four different levels. The first observation was that noise decreases performance on individual retrievals, such that additional attempts are required to achieve the same number of successful retrievals. The second observation is that as noise increases, the number of attempts required to achieve comparable performance to human difficulty also increases. This is expected, but what is surprising is that the quality of the fits as noise and number attempts increase together are comparable (or sometimes better) to the results without noise. Thus, this provides us with two parameters that appear to trade off: noise and attempts.

7.3 Process Model

A knowledge base and retrieval algorithms are two parts of our overall model implemented in Soar. That model first retrieves each of the prompt words into working memory, and those words become sources of activation spread in long-term semantic memory. The best model then attempts to retrieve a word, and that retrieval is biased by the spread activation and noise. Once a word is retrieved, it is tested to ensure that there exists association links to all the prompt words. If all of those links do not exist, the model inhibits the retrieved word and tries again, repeating this process until it finds an answer or runs out of attempts. If it runs out of attempts, it selects the retrieved word with the most associations to prompt words, breaking ties randomly.

The main variation in the model is the number of attempts used to find a potential solution. We performed a sweep to find the number of attempts that led to the best match to human data. That best attempt number for 2 seconds is 1, which is always less than the number for 7 seconds, which is also always less than the number for 15 seconds, except in one case where they were equal. Across the two best knowledge bases (SWOWEN and HBC) without noise, there was a spread from 2 (HBC) to

5 (SWOWEN) retrievals for 7 seconds, and 3 (HBC) to 8 (SWOWEN) retrievals for 15 seconds. With noise, as mentioned earlier, the number of attempts grew, raising the number of attempts to 4 and 6 for HBC and 8 and 12 for SWOWEN. Given the overall high quality of the fits across the range of noise and attempts, it is prudent to avoid speculation on exact levels of noise and attempts from these data alone. It is also difficult to determine these values using the temporal dynamics available when modeling tasks in cognitive architectures such as Soar and ACT-R for a simple model such as ours that does not include reading, eye movements, motor actions, etc. However, the main result is that our model predicts that as time increases, more retrievals are possible, which increases the number of correct responds, which is in accordance with human performance.

7.4 Best Model Summary

Figure 5 provides a more detailed plot of the best model variant, which uses the SWOWEN knowledge base and 0.5 noise, modeling 15 second human behavior. It has an extremely high R^2 of 0.98 and low MSE of 0.29. Each data point is one of the twelve groupings of twelve questions, with the y-axis being the number of RAT question that were answered correctly by the model (averaged over 100 runs), and the x-axis being the number of questions correctly answered by humans (averaged over 289 subjects).

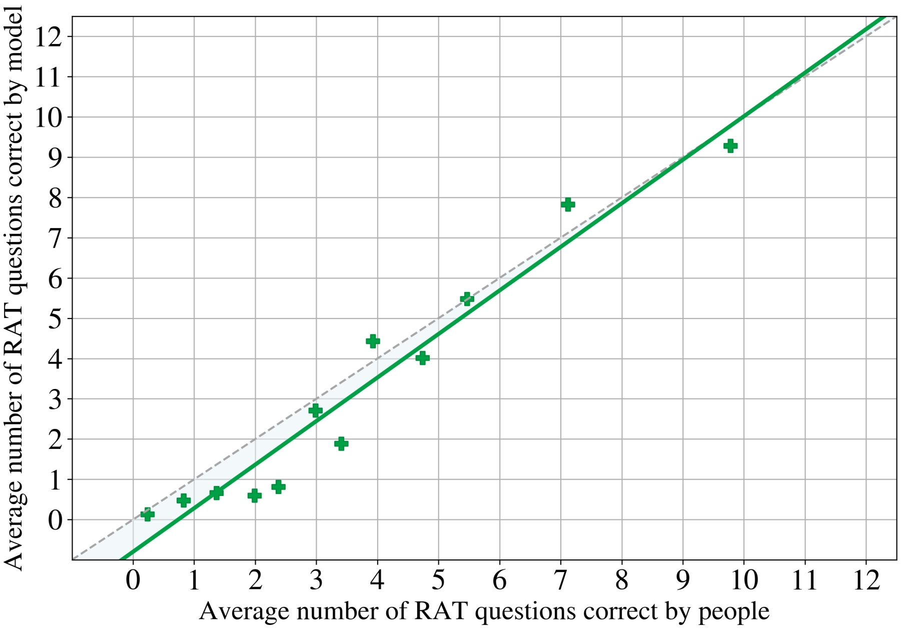

Fig. 5. Performance of the best model variant (SWOWEN, 0.5 noise, 15 seconds) plotted versus human performance in terms of problem difficulty.

8 Conclusions and Future Work

We derive three major conclusions from this work. First, that the many variations explored in this work, including base-level activation, noise, and knowledge sources, show the value of comprehensive modeling. It leads the way for future modeling work, outside of the RAT, to include results from adjusting various parameters that are normally preset and fixed. For example, our work found a trade off between attempts given to the model and noise added. In other work, noise is usually a fixed parameter. Exploring these parameters gives new insight into the robustness of our results by showing that small variations in parameters such as the number of attempts and noise do not lead to a large change in the results.

Our second conclusion is that this work provides confirmatory evidence of the most important factors influencing retrieval from long-term semantic memory. The performance of the best models is striking, with high R^2 and low MSE. The results suggest that people iteratively retrieve information from long-term semantic memory using spreading activation biased by directional association strengths. This result is robust even with variations in noise and base-level activation and provides evidence in favor of the current theory but with additional detail.

The final conclusion relates to how people store and or retrieve word association knowledge. This work examined five potential sources of word association knowledge that the model could draw from. We found that the models that best matched human behavior use a large knowledge base of free associations collected directly from people. This is in contrast to using knowledge from a large English corpora such as direct word co-occurrence frequency or a distributional language model such as word2vec or GloVe. Thus, we conclude that people's semantic word association knowledge, or at least what they use in this task, differs from word frequency data extracted from written text. This work raises the question as to how people differentially spread activation through only those associations relevant to the RAT.

To expand on the current work it would be informative to adapt and evaluate this model to RAT problems that require different types of associations. The problems used in this paper, while popular, only examine compound associations. Other RAT problems require functional associations such between "man" and "uncle" (Worthen & Clark, 1971). Examining this could help evaluate the robustness of the model and further determine what knowledge humans store in their memory that they recall for word association tasks.

A second direction for future work would be to attempt to model individual behavior instead of the summary behavior of groups of individuals. We assume that there is variation between participants on a task such as the RAT, but it is not clear whether variances in the factors we have identified explain those differences. Individual task performance data, especially if it is augmented intermediate answers would make possible a more detailed evaluation of our model than was possible using the data modeled here. Addition bio-metric data, such as eye movements, would allow us to develop more complete and detailed process models that predict the number of attempts directly from the timing predictions of the model instead of treating them as a parameters. Moreover, these data would allows us to explore whether more complex strategies are better fits to human performance than our simple iterative model.

The final direction for future work we see is applying this model or parts of this model to other semantic retrieval tasks, such as word-sense disambiguation in natural language process. This could lead to further insights into the retrieval mechanisms at play in tasks such as the RAT.

9 Acknowledgments

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474.

Anderson, J. R. (1990). *The adaptive character of thought.* Psychology Press.

Anderson, J. R., & Pirolli, P. L. (1984). Spread of activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 791.

Benedek, M., & Neubauer, A. (2013, 12). Revisiting Mednick's model on creativity-related differences in associative hierarchies. Evidence for a common path to uncommon thought. *The Journal of Creative Behavior*, *47*, 273-289. doi: 10.1002/jocb.35

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

Davelaar, E. J. (2015). Semantic search in the remote associates test. *Topics in cognitive science*, *7*(3), 494–512.

Davies, M. (2008). *The Corpus of Contemporary American English.* BYE, Brigham Young University.

De Deyne, S., Navarro, D., Perfors, A., Brysbaert, M., & Storms, G. (2018, 10). The âsmall world of wordsâ english word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*. doi: 10.3758/s13428-018-1115-7

Derbinsky, N., & Laird, J. E. (2010). Extending Soar with dissociated symbolic memories. In *Symposium on Human Memory for Artificial Agents, AISB* (pp. 31–37).

Gabler, K. (2013). *Human Brain Cloud.* Retrieved from `https://humanbraincloud.com/`

Jones, S. J., Wandzel, A. R., & Laird, J. E. (2016). Efficient computation of spreading activation using lazy evaluation. In *Proceedings of the 14th International Conference on Cognitive Modeling.*

Kajić, I., Gosmann, J., Stewart, T. C., Wennekers, T., & Eliasmith, C. (2016). Towards a cognitively realistic representation of word associations. In *Cogsci.*

Kajić, I., Gosmann, J., Stewart, T. C., Wennekers, T., & Eliasmith, C. (2017). A spiking neuron model of word associations for the remote associates test. *Frontiers in psychology*, *8*, 99.

Kajić, I., & Wennekers, T. (2015). Neural network model of semantic processing in the remote associates test. In *Coco@ nips.*

Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*.

Lebiere, C., & Best, B. J. (2009). Balancing long-term reinforcement and short-term inhibition. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2378–2383).

Lee Bae, C., Huggins-Manley, A., & Therriault, D. (2014). A measure of creativity or intelligence? Examining internal and external structure validity evidence of the Remote Associates Test. *Psychology of Aesthetics Creativity and the Arts*, *8*. doi: 10.1037/a0036773

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus.

Marko, M., Michalko, D., & Riecansky, I. (2019, 12). Remote Associates Test: An empirical proof of concept. *Behavior Research Methods*, *51*, 2700–2711. doi: 10.3758/s13428-018-1131-7

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220-232.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moss, J. (2006). The role of open goals in acquiring problem relevant information. *Unpublished doctoral dissertation, Carnegie Mellon University*.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* http://www.usf.edu/FreeAssociation/.

Olteţeanu, A.-M., & Falomir, Z. (2015). ComRAT-C - A computational compound Remote Associates Test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, *67*, 81–90.

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*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Retrieved from http://www.aclweb.org/anthology/D14-1162

Smith, K., Huber, D., & Vul, E. (2013, 04). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*, 64-75. doi: 10.1016/j.cognition.2013.03.001

Taatgen, N. A., Lebiere, C., & Anderson, J. R. (2006). Modeling paradigms in act-r. *Cognition and multi-agent interaction: From cognitive modeling to social simulation*, 29–52.

Valba, O., Gorsky, A., Nechaev, S., & Tamm, M. (2021). Analysis of english free association network reveals mechanisms of efficient solution of remote association tests. *PloS one*, *16*(4), e0248986.

Worthen, B. R., & Clark, P. M. (1971). Toward an improved measure of remote associational ability.

Journal of Educational Measurement, *8*(2), 113–123.

Wu, C. L., Huang, S. Y., Chen, P. Z., & Chen, H. C. (2020). A systematic review of creativity-related studies applying the Remote Associates Test from 2000 to 2019. *Frontiers in Psychology*, *11*(573432).

A Model Parameters

In Soar, there are several parameters to control semantic memory. These parameters control both base-level activation and spreading activation calculations. For brevity, only the parameters that have been set to non-default values are listed. They are all set using Soar's command line interface using commands of the form "`smem --set [parameter-name] [parameter-value]`." The commands are described here, then the specific values for the parameters used in each model are listed. Activation noise is not part of the official Soar release, but was adjusted similarly.

The following list of parameters control the base-level activation calculation: `activation-mode`, `base-update-policy`, `base-incremental-threshes`, and `base-inhibition`. `activation-mode` is used to turn on base-level activation and the parameter value for this is simply `base-level`. `base-update-policy` determines the timing by which base-level activation values are updated. Because it is expensive to recompute base-level activation values for the entire database of stored elements, an approximation is used. This approximation is used by setting the `base-update-policy` parameter to `incremental`. `incremental` updates only a portion of the base-level activation values depending on when they were last boosted. The parameter that controls the update times is `base-incremental-threshes`. Because base-level activation features short-term transient change followed by a long-term decay, we select values for the `base-incremental-threshes` parameter that update elements more frequently when they have recently been accessed, but then recompute less often as the decay becomes the dominant behavior for the value. The values we use for `base-incremental-threshes` to do this are "`1, 2, 4, 8, 20`." We also tested with updating of the entire database at every query to ensure results did not significantly differ by setting the parameter `base-update-policy` to `naive`. `base-inhibition` is used to toggle usage of an additional dynamics to base-level activation beyond the default of a transient peak, followed by a decay. Instead, there is a short-term inhibition in a recently-boosted element consistent with the implementation specified by Lebiere and Best (2009). This is used by setting the `base-inhibition` parameter to `on`.

The parameters that control the spreading activation calculation are `spreading`, `spreading-depth-limit`, and `spreading-limit`. `spreading` is set to `on` to enable spreading activation calculations. `spreading-depth-limit` specifies the limit to the depth of spread in the semantic memory network. For all models, this value was set to `1` because of the nature of the HBC data.`spreading-limit` is used to limit the total number of elements that a given source element can spread to. It is primarily used as a parameter to limit computational cost. Because these models ran quickly in 64.278 msec, we set this value to `9999999`, which was never reached with depth `1` spread. (The highest fan prompt word in the network spread to 765 elements.)

While we tested several variants, the versions the model used for most of our analysis used the same settings for activation calculation.

Table 16

Semantic memory activation parameters used in most models.

Parameter	Value
activation-mode	base-level
base-update-policy	incremental
base-incremental-threshes	1 2 4 8 20
base-inhibition	on
spreading-depth-limit	1
spreading-limit	9999999

B Statistical Analysis

We make several simplifying assumptions to support a statistical analysis. In much of the literature on human RAT performance, there is a notion that independent from considering specific humans' situations, you can conceive of a RAT problem as having an "inherent" difficulty. We formalize this approximation of diverse individual humans' aggregate performance as the foundation of our analysis. Consider each of the 144 RAT problems as each having a distinct and independent probability p_i (where $1 \leq i \leq 144$) of being answered correctly. For a given question, that question's probability of being answered correctly is the same for each human and independent from other questions. Obviously this is not precisely true, but it allows a simple statistical analysis.

Using this assumption, we can treat the probability distribution for how many people will answer a given question i correctly as a binomial distribution that depends on the underlying probability p_i and the number of people, 85 (in the 7 second condition). We do not know what p_i is. But, using a beta distribution as a conjugate prior, we can predict how similar experiments would proceed using a beta-binomial distribution. We use Jeffrey's prior ($\alpha = \beta = \frac{1}{2}$). (A uniform prior gives similar results.) Thus, for question 1 in which 71/85 people answered correctly, we believe the probability that n people would answer question 1 correctly in another experiment as $P(n) = \mathbf{BetaBin}(n|\alpha = 71 + \frac{1}{2}, \beta = 85 - 71 + \frac{1}{2})$.

Sampling from that distribution provides additional simulated results on that question. We can do this for each of the 144 questions to create a complete simulated experiment.

The purpose of doing this sampling is to generate a statistic by which we can claim there is a significant difference between our models. Given this simulated experiment data, we can record a new MSE for our models. Thus, to provide a claim that one model is significantly better than another, we sample from all 144 questions' inferred beta-binomial distributions 10000 times to create 10000 simulated experiments for the 144 problems. Then, we order each model by MSE 10000 times. We use the proportion of the experiments in which a relative ranking by MSE changed between two models as the likelihood that ordering by MSE could occur by chance. We interpret that likelihood as a p-value.