REM (1997) Predicts Recognition Tested With 2AFC, and 4AWC

Zainab Rajab Mohamed (zrmohame@iu.edu)
Department of Psychological and Brain Sciences
Indiana University, Bloomington, IN, USA

Constantin G. Meyer-Grant (constantin.meyer-grant@psychologie.uni-freiburg.de)
University of Freiburg, Freiburg, Germany

Richard Shiffrin (shiffrin@indiana.edu)
Department of Psychological and Brain Sciences
Indiana University, Bloomington, IN, USA

Abstract

We use a novel paradigm to test models of long-term recognition memory: After studying lists, tests are made with two items, both OLD, both NEW, or one of each. Some tests used Two-Alternative Forced Choice (2AFC) in which Ss were asked to choose the item more likely OLD (Experiment 2 asked Ss to choose the item more likely NEW); other tests used four-way classification (4WC) in which Ss were asked to classify the two items as 1) both old, 2) both new, 3) left old, right new, or 4) left new, right old. Each S studied lists containing 12 words, 24 words, 12 pictures, 24 pictures, or lists of 12 words randomly mixed with 12 pictures (so tests were both words, both pictures or one each). All the choice probabilities were predicted well by the Retrieving Effectively from Memory model (REM) of Shiffrin and Steyvers (1997) using mostly the three 1997 parameter values and the REM decision threshold of odds of 1.0. Signal-detection modeling (unequal variance Gaussian strength distributions) predicted the choice probabilities with different parameters for different conditions. Initial analysis and modeling of Response times suggested that REM may be well suited to predict differing accuracy and response time results for judgments of OLD and NEW.

Keywords: recognition memory; forced choice; memory modeling; REM; old vs. new judgments

Introduction

This research explores the processes of recognition memory, and assesses models of recognition and their ability to generalize to new tasks. Most recognition studies in the laboratory present participants with a list of items, usually words or pictures, and then test memory by presenting single items one at a time for a judgment whether or not the test item had been studied (a 'target' or OLD) or no studied (a 'foil', or NEW). We use a novel paradigm in which each test is of two items, side by side, equally often two targets, two foils, or left one old, right one new, or left one new, right one old (see also Voormann et al. 2021). After some lists participants are given a two alternative forced choice task (2AFC) and asked to choose the item more likely old. Note that there is no correct answer when tests are of two targets or two foils; why we test such cases for 2AFC will become clear when we discuss certain response time results. After other lists participants are asked to carry out a four -way classification (4WC), requiring them to classify the two test items using four keys: both targets, left one a target, right one a target, both foils. Lists have 12 words, 24 words, 12 pictures, 24 pictures, and mixtures of 12 words and 12 pictures. For mixed lists, some tests have two words, other two pictures, and others one word and one picture. All these conditions are used for every participant. Both choice and response times (RT) are measured. For 2AFC, some tests have no correct answer. For 4WC, on the other hand, every trial has one correct answer and three distinguishable incorrect answers. Exp. 1 asked participants to judge oldness (which item was more likely old), while Exp. 2 asked for newness judgments (which item was more likely new). Additional design details will be given later in this article.

There are over 120 conditions with distinguishable results, each with a choice and a response time. It is hard to keep track of so many results without a model to make sense of them, so we next present the first model we applied, the Retrieving Effectively from Memory model (REM) of Shiffrin and Steyvers (1997, 1998).

REM Model

REM was a very simple (in fact oversimplified) three parameter Bayesian inspired model developed to predict accuracy of decisions for recognition tasks using single item tests for target vs foil decisions. Simple as it was, it predicted the patterns of results found in a wide variety of tasks. Each item in REM is represented as a vector of 20 feature values, each value selected randomly from a geometric distribution (with parameter g) representing environmental base rates for this class of item. When study time is t, the probability that a value will be stored for a given feature is u. If stored, the value is stored correctly with probability $\boldsymbol{c}$. Otherwise (with probability 1-c), the value stored is a random choice from the base rate distribution (again with parameter g).

When an item is tested, its 20 feature values are compared in parallel to each of the stored traces. The comparison produces a likelihood ratio for each feature; these are multiplied to produce a likelihood ratio for each trace. The average likelihood ratio gives the odds that the trace was stored due to study of the test item vs stored due to study of some other item (either another study item, or an item never studied; see Eqs. 4 A, B in the 1997 article). It is optimal to respond 'target', or OLD, if the odds value is greater than 1.0.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

This decision is optimal in the sense that it uses all the information available in the stored traces and the test item.

We adapted REM for two item tests in the simplest and most straightforward way: An odds value is calculated for each item. For 2AFC the choice is the item with the largest odds. For 4WC each item is judged to be a target (i.e. OLD) if its odds value is greater than 1.0. That produces two decisions, each OLD or NEW, and they are used to give one of the four responses. REM was developed to predict choice accuracy, not response time, so the predictions we give and discuss in this article focus on accuracy. That said, a few significant response time results for which REM might give plausible accounts will be presented and discussed as well.

The pictures and words we used had often been studied and it was known that pictures were better recognized than words (indeed that is what we found). Therefore, to apply REM to our data, we let u and g match the values used in the 1997 article (u =0.04; g =0.4) and let c have two values, c_p slightly higher for pictures, and c_w slightly lower for words (c_p =0.75; c_w =0.55). A final additional assumption was made that there is zero similarity between pictures and words, so that in mixed lists only same category traces are compared.

Results and Predictions

REM model predictions

Figure 1 gives the results and predictions for the 2AFC conditions testing words or pictures, only one of which is a target, so have correct answers. Pictures have higher performance than words, predicted due to the higher c value for pictures. Performance is better for shorter lists, predicted because the distributions of odds has higher variance and skewness for longer lists, the extra noise reducing performance (as will be discussed in more detail later).

Figure 2 gives the results and predictions for the 2AFC conditions testing one word and one picture. These results include conditions with both targets and both foils, because there are strong biases in the choices. When just one item is a target (OLD), performance is very good and about equal when the picture or the word is the target. However, there is a huge bias to choose the picture when both are targets and to choose the word when both are foils. REM predicts this bias due to the better encoding of pictures: Pictures tend to have higher odds than words when both are targets and tend to have lower odds than words when both are foils.

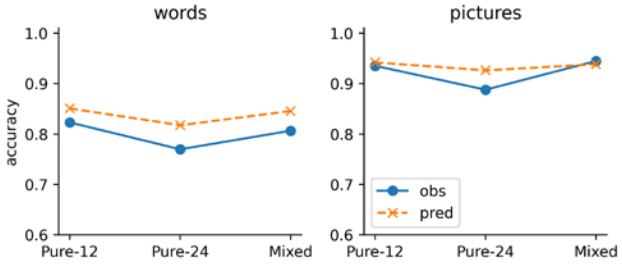

Figure 2. 2AFC: Observed and predicted accuracy for pure tests with one old and one new item tested -- two words after study of 12 or 24 words; two pictures after study of 12 or 24 pictures; or mixed: two words or two pictures tested after study of a mixed list of 12 words and 12 pictures.

Figure 3 gives the results and predictions (as dots) for the 4WC conditions. In each panel there is one correct response and prediction and three incorrect responses and predictions. The predictions arise naturally from the tendency for picture targets to have higher odds than words, and picture foils to have lower odds than words.

Discussion: Given we used the simplest 1997 REM model unchanged except for better learning for pictures than words, (the c values), it was surprising to us that all the trends in the data were predicted correctly, and that the predictions were reasonably close quantitatively. We believe it is noteworthy that the predictions for every condition in the 1997 article and in the present data used the same REM equation to produce the odds, and used the same decision criterion, odds of 1.0. How is it possible to use the same decision criterion? That arises from the normative Bayesian derivation of REM and the positions and shapes of the odds distributions. The odds distributions are very skewed and long-tailed, especially for targets. We will say more about this later in the article. The fact that REM was developed for single item recognition tests and generalized so well to predict 2AFC and 4WC is one measure by which models are evaluated. It seems clear that REM passed this 'test' with flying colors.

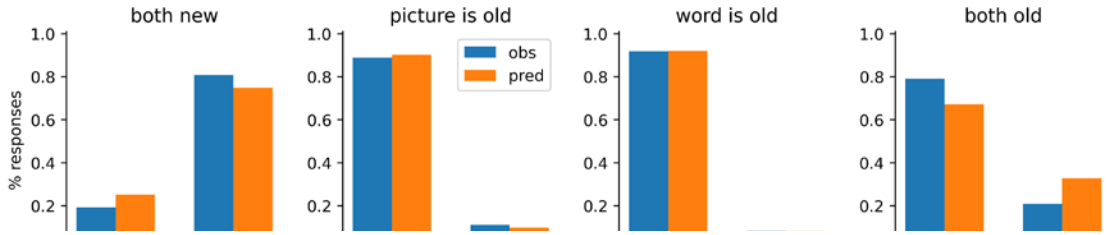

Figure 1. 2AFC: Observed and predicted responses for mixed tests with one word and one picture. Test conditions are given by labels above each panel, and choice responses below each bar. Both-new and both-old have no correct answer, but the choices of word vs picture differ.

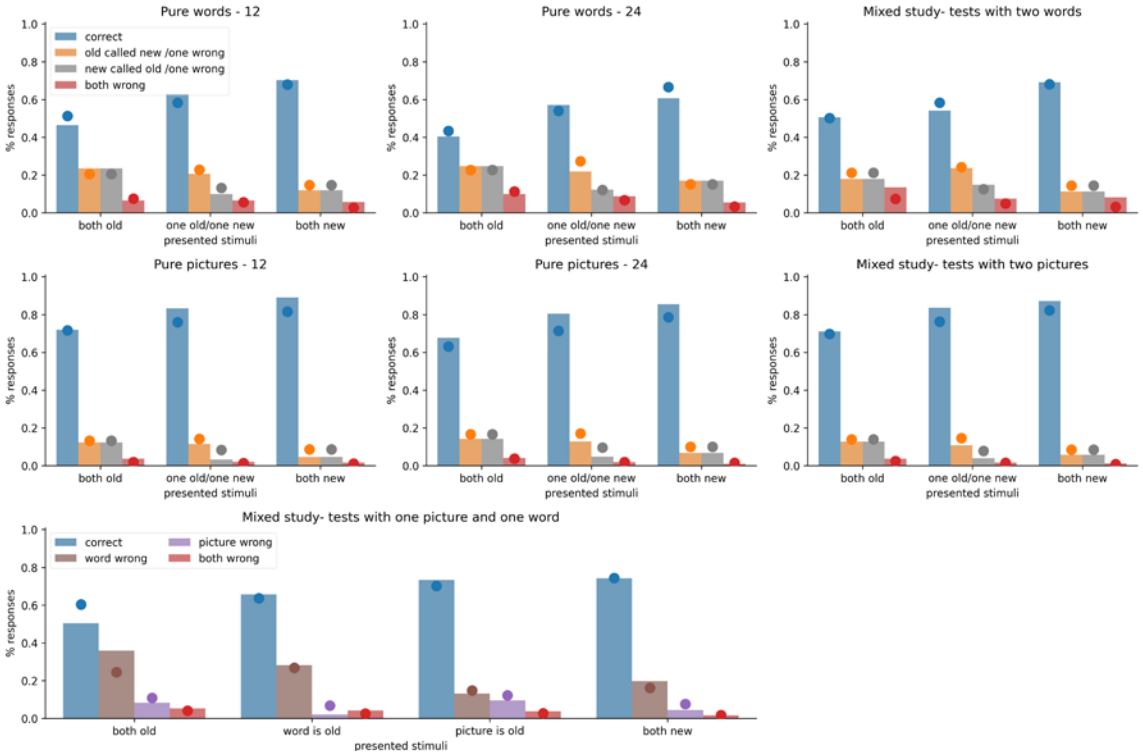

Figure 3. 4WC: Observed data as bars, and predictions as dots. The study conditions are given above each pane, and the test presentation conditions given by the labels below each group of four. The four possible responses are given by the four bars, with correct responses always given by the blue bar on the left, and doubly incorrect responses given by the red bar on the right. Singly incorrect responses are given by the two central bars -- see the description in the main text.

Signal detection model predictions

It has been traditional to describe recognition results by assuming that foils have a Gaussian distribution of strengths with a mean and variance that is higher for targets. It is also typical to allow different means and perhaps variances for different conditions. We also fit such signal detection models to the present data. The predictions (not shown to save space) were quite good but as expected it was necessary to allow different parameter values for different conditions ($\mu_{\mathrm{new/w}} = 0$, $\sigma_{\mathrm{new/w}} = 1$, $\mu_{\mathrm{old/w/12}} = 1.67$, $\mu_{\mathrm{old/w/24}} = 1.60$, $\sigma_{\mathrm{old/w}} = 1.80$, $\mu_{\mathrm{new/p}} = -1.64$, $\sigma_{\mathrm{new/p}} = 1.61$, $\mu_{\mathrm{old/p/12}} = 6.39$, $\mu_{\mathrm{old/p/24}} = 5.34$, $\sigma_{\mathrm{old/p}} = 5.15$; the location of the criterion at 0.95 was estimated from the data as well). An important observation is that this approach requires the distributions of new pictures and new words to differ from each other. Otherwise, the model is unable to predict the stark preference for choosing a word in 2AFC when both a new picture and a new word are presented together. This problem can be resolved by implementing a decision rule based on likelihood-ratios (and a fixed criterion of 1.0) instead of raw memory-strength signals (see, e.g., Glanzer et al. 2009; Osth et al., 2017). Interestingly, this happens to be the same decision rule as the one used by REM. However, such modeling is descriptive, in contrast to REM which provides causal processes to account for the results.

A Noteworthy Response Time Finding: Figure 4A gives median response times for 2AFC conditions without correct answers when judging old: Judging the more OLD item of two targets takes far less time than judging the more OLD of two foils (overall 1186 ms vs 1676 ms). Suppose that that the time to choose the better item is determined by the difficulty of judgment, and that difficulty is indexed by the difference of the two odds. Then judgments would be faster when the two odds are very different in magnitude, and slower when close in magnitude. The higher variance for target distributions means that two samples from that distribution will tend to be more distant in value than two samples from the lower variance foil distribution.

REM Odds Distributions

We next illustrate the larger odds difference for two targets than for two foils. The odds distributions in REM are so skewed, and long tailed, especially for targets, they cannot be graphed in a meaningful way. Thus Figure 5A shows the distributions of the target and foil odds raised to the 1/11th power for lists of 12 words and 12 pictures (This monotonic transform does not change the accuracy predictions).

If one should take two samples from the target distribution they would differ more on average than two samples from the foil distribution. This is illustrated in Figure 5B showing the difference distribution for 12 words. If one should map these differences into median RT with a suitable inverse monotonic transform it is clear that two targets would be much faster.

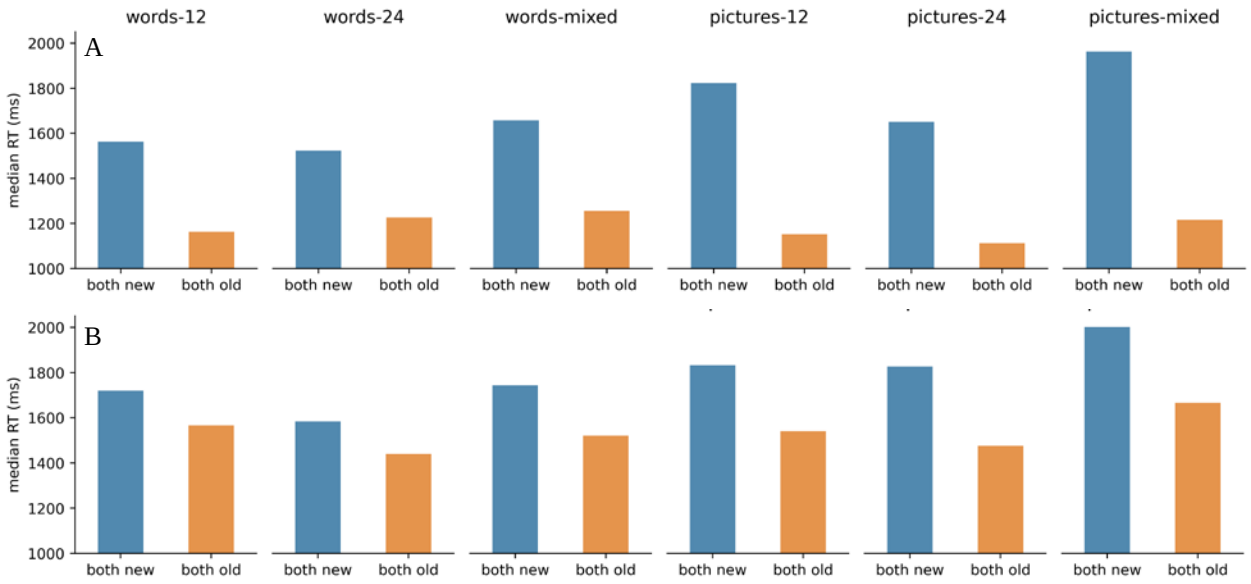

Figure 4. 2AFC: Observed median response times for conditions without correct answer for experiment 1 (oldness judgment) in panel A, and experiment 2 (newness judgment) in panel B. Test conditions are given by labels above each panel, and choice responses below each bar.

Is it reasonable that the major determiner of median RT in 2AFC is the difference in odds. Figure 5C shows for correct responses in various 2AFC conditions the observed median RT (dots) and predictions (stars). The predictions are based on a monotonic transform of the absolute difference in REM 1/11th odds for each condition, which is given by:

$$RT = \boldsymbol{\alpha} + \boldsymbol{\beta}|\Phi_{diff}|^{\delta},$$

where $\boldsymbol{\alpha}$ is the base time, $\boldsymbol{\beta}$ is a scaling parameter, Φ are 1/11th REM odds and $\boldsymbol{\delta}$ is the power. The predictions in 5C are based on, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ equal to: 500, 864 and -0.14 respectively.

Judging Old vs. Judging New: Experiment 2

Brainerd, Bialer, Chang, and Upadhyay (2021) and Meyer-Grant and Klauer (2023) both reported data showing that the instructions to judge whether a test item is old produces a different pattern of results than instructions to judge whether a test item is new. We therefore carried out an experiment identical to Exp. 1, except with 2AFC instructions to choose the test item more likely to be new. Space does not permit reporting the results, but the accuracy findings were generally similar to those from Exp. 1. Operating within the REM framework, it seemed reasonable to us that a participant asked to choose a new item would use the odds of new, which would be 1/(odds of old). That is, if 'judge old' instructions would lead a participant to calculate for each test item the odds of old, Φ, then 'judge new' instructions would lead the participant to calculate for each item the odds of new, 1/Φ. For 2AFC the choice of the new item would be the greater of 1/Φ, and for 4WC each item would be judged new if 1/Φ were above 1.0. Space in this report only allows us to report that this REM model with a slightly lower $\boldsymbol{c}$, and a slightly adjusted decision criterion of 1.2 rather than 1.0 for words did a good job of predicting the accuracy results of Exp. 2, including the small changes from Exp. 1.

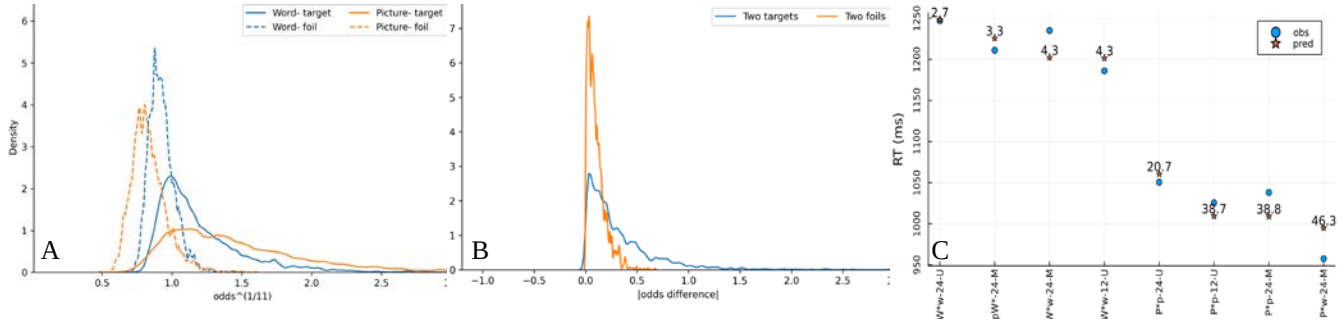

Figure 5. A: distribution of REM odds for words and pictures, targets and foils. B: distribution of odds difference between two targets and two foils for 12 word lists. C: Observed and predicted median response times for correct responses in 2AFC. Conditions are given on the x-axis, with stars representing the old item (whether picture or word), 12/24 indicates list length, and "U/M" represents pure/mixed lists respectively. Numbers above each dot gives the odds difference predicted by REM.

Here we want to report a remarkable response time finding from Exp. 2. In Exp. 1 we saw a much slower 2AFC judgment for two new items (foils) compared to two old items (targets). When judging new, one might expect this result to reverse. In fact, this is the prediction for REM modeling using the raw odds and 1/odds, and for analogous versions of signal detection models that use likelihood ratios for a decision. The results shown in Figure 4B showed the opposite result: two new items were slower than two old items, albeit the differences were smaller than in Exp. 1 (Figure 4A).

This unexpected result is predicted by REM if decisions (for both judgments of old and new) are based not on the raw odds, but on a highly compressed transform, such as the odds or 1/odds raised to the $1/11^{th}$ power, as shown by the REM difference distributions in Figure 6 for 12 words.

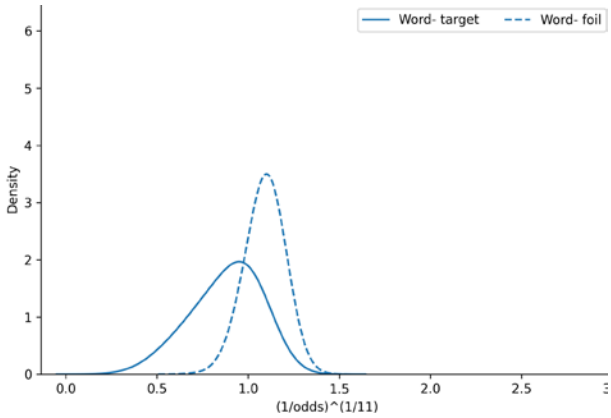

Figure 6. Distribution of REM inverse odds for 12-word lists for targets and foils.

These difference distributions have the right character, because for both odds and 1/odds the difference between two samples from the old distribution tends to be larger than the difference between two samples from the new (the same applies to signal detection modeling if decisions are based on ratios rather than raw strengths).

That decisions should be based on a monotonic compression of the raw odds is reasonable: In REM the distributions of raw strengths are so long tailed it seems implausible that that those values could be used directly for decision making. Such monotonic compression will leave the accuracy predictions unchanged: If the raw odds for one item are larger than for another, so will be the compressed odds; if an item is judged old because the raw odds is greater than 1.0 (or judged new because the raw 1/odds is greater than 1.0) than so will be the judgments based on compressed odds.

Discussion

The very simple and very incomplete three parameter REM model of Shiffrin and Steyvers (1997, 1998) seemed to have captured enough of the important processes of recognition memory to predict several of the main trends observed in studies of choice accuracy using single word OLD/NEW tests. We have seen that the same model captures the main findings in two item testing, using both words, pictures, and mixtures of words and pictures. Every condition in the studies using single item tests in 1997, and in the present studies using two item tests were modeled with the same REM equation (Eqs. 4 A,B in the 1997 article) and the same decision criterion of odds of 1.0. That the 1.0 criterion can be used in every condition is due to the positions and the extremely skewed shapes of the odds distributions predicted by REM. That a compression of the odds is the appropriate decision statistic is suggested by the correct predictions of the direction of response time differences for 2AFC tests without correct answers, for both judgments of old and judgments of new.

The successes of this simplified REM model notwithstanding, we must point out that it omits several components that should be part of any recognition model, including at least the following four. Some of these, but not all, were discussed and modeled in the 1997 and 1998 articles.

1. Context features: The features representing items stored in traces, and used to probe memory, should include context features. Some context features might change within list, between study and test, between lists, and during delays between study and test (see Mensink & Raaimakers (1988) for one application). Other context features might remain constant during these times. In addition it seems likely that participants might be able to choose which context features to include at these times, though there is evidence that there are 'default tendencies' that occur at study: Malmberg and Shiffrin (2005) provided evidence that context tends to be stored only in the first few seconds of study, the rest of the study time presumably devoted to storage of content such as coding of associations. It is also likely that participants are able to 'reconstruct' context during retrieval in certain conditions (e.g. Bauml & Trissl, 2022; Krichbaum & Bauml, 2023).

2. Associative features: These should be among those stored and used at test. These include associations among items on a study list. Such features are related to those at the heart of the Temporal Context Model (TCM) of Howard and Kahana (2001) and its successors.

3. Storage during testing: Traces are stored during testing, accounting for example for output interference (e.g. Criss, Malmberg & Shiffrin, 2011; Annis, Malmberg, Criss & Shiffrin, 2013). In addition, retrieval of stored traces matching a test trace in many cases will cause strengthening of the retrieved trace rather than storage of a new trace, important to account for differentiation and the list strength effect (see Ratcliff, Clark & Shiffrin, 1990 and Shiffrin, Ratcliff & Clark, 1990).

4. Recall during recognition: REM is what is often termed a 'familiarity' model, in which a summed activation of memory traces produces a sense of familiarity used to make a recognition decision. There is evidence that in many standard recognition paradigms familiarity is sufficient to produce accurate predictions (see Dunn, 2008, and witness the present application of REM as a prime example).

Sufficiency does not preclude recall or recollection and as noted in Shiffrin and Steyvers (1997), strengthening a retrieved trace during study or test suggests recall of that trace. It seems likely that recall does occur during recognition testing even if familiarity is used to decide. Recall during recognition was a prime motivation for a large number of studies that have asked participants who make old recognition decisions to classify them as 'remember' (i.e. recall) or 'know' (i.e. familiarity) – see Gardiner and Java (1990).

Future directions

REM was derived to predict recognition accuracy (and implicitly choice), and has been applied here to predict choice accuracy in a set of new recognition paradigms. Some intriguing response time results from 2AFC were shown to be consistent with a likely extension of REM to predict response times as well, but a model that might predict both choice and response times for our studies remains a challenge for future research.

Method

Stimulus material

The experiment included 384 words and 384 pictures. Words with medium (20-100 fpm) frequency were selected from the *SUBTLEX-UK* (Van Heuven et al., 2014) word dataset, and the pictures were randomly drawn from Brady et al. (2008) image dataset that included images from distinct categories such as animals, plants, etc.

Participants

Experiment 1: 83 undergraduate students from Indiana University Bloomington subject pool participated in the experiment as part of a course requirement. Sessions lasted for around 45 minutes, and participants were debriefed after the experiment. Experiment 2: 86 undergraduate students.

Design

Participants started with practice of both 2AFC and 4WC to familiarize them with the experiment. Then each participant was tested in every condition, receiving one block each of 2AFC and 4WC (order randomized). Each block had eight lists followed by tests. For a given participant no stimulus was studied or tested more than once. Each list was followed by simple arithmetic operations, summing a series of single digits (taking about 15 seconds) to clear short-term memory. The experiments were coded with Psychtoolbox on Matlab.

References

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1365–1376. https://doi.org/10.1037/a0032188

Bäuml, K. T., & Trißl, L. (2022). Selective memory retrieval can revive forgotten memories. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(8), e2114377119. https://doi.org/10.1073/pnas.2114377119

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Brainerd, C. J., Bialer, D. M., Chang, M., & Upadhyay, P. (2022). A fundamental asymmetry in human memory: Old ≠ not-new and new ≠ not-old. *Journal of experimental psychology. Learning, memory, and cognition*, *48*(12), 1850–1867. https://doi.org/10.1037/xlm0001101

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*(4), 316–326. https://doi.org/10.1016/j.jml.2011.02.003

Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, *115*(2), 426–446. https://doi.org/10.1037/0033-295X.115.2.426

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*(1), 23–30. https://doi.org/10.3758/BF03202642

Glanzer, M., Hilford, A. & Maloney, L.T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review, 16*, 431–455. https://doi.org/10.3758/PBR.16.3.431

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46, 269–299.

Kriechbaum, V.M., Bäuml, KH.T. (2023). The critical importance of timing of retrieval practice for the fate of nonretrieved memories. *Sci Rep 13*(6128). https://doi.org/10.1038/s41598-023-32916-7

Malmberg, K. J., & Shiffrin, R. M. (2005). The "One-Shot" Hypothesis for Context Storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 322–336. https://doi.org/10.1037/0278-7393.31.2.322

Mensink, G.-J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review, 95*(4), 434–455. https://doi.org/10.1037/0033-295X.95.4.434

Meyer-Grant, C. G. & Klauer, K. C. (2023). Does ROC asymmetry reverse when detecting new stimuli? Reinvestigating whether the retrievability of mnemonic information is task-dependent. *Memory & Cognition, 51*(1). 160–174. https://doi.org/10.3758/s13421-022-01346-7

Osth A. F., Dennis S. J., Heathcote A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, 92, 101–126.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 163–178. https://doi.org/10.1037/0278-7393.16.2.163

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 179–195. https://doi.org/10.1037/0278-7393.16.2.179

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. *Rational models of cognition*, 73-95.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Voormann, A., Spektor, M.S. & Klauer, K.C. (2021). The simultaneous recognition of multiple words: A process analysis. *Mem Cogn* 49, 787–802. https://doi.org/10.3758/s13421-020-01082-w