Does Memory Activation Grow with List Strength and/or Length?

David E. Huber, Heidi E. Ziemer, and Richard M. Shiffrin

Indiana University
Bloomington, IN 47405
dhuber@ucs.indiana.edu
hziemer@ucs.indiana.edu
shiffrin@ucs.indiana.edu

Kim Marinelli

Northwestern University
Evanston, IL 60208
kim@nugm.psych.nwu.edu

Abstract

Recognition of an item from a list is typically modeled by assuming that the representations of the items are activated in parallel and combined or summed into a single measure (sometimes termed 'familiarity' or 'degree-of-match') on which a recognition decision is based. The present research asks whether extra items (length), or extra repetitions (strength), increase this activation measure. Activation was assessed through examining hits and false alarms as the length or strength of word categories were varied. The use of a categorized list insured that response criteria were not changed across the length and strength manipulations. The results demon-strated that: 1) The activation does not change with an increase in the strength of presented items other than the test item; and 2) The activation is increased by an increase in the number of presented items in a category. The results provide important constraints for models of memory, because most models predict or assume either that activation grows with both length and strength, or grows with neither. In fact, the only extant model that can predict both the length and strength findings is the differentiation version of the SAM model (Shiffrin, Ratcliff, & Clark, 1990).

Introduction

The explosion of interest in memory models associated with the advent of neural net and connectionist frameworks has called into focus certain fundamental assumptions about memory representation and retrieval. Many models assume that the result of probing memory is the activation of all of memory. Especially for recognition tests, it is typically assumed that the recognition decision ('old' or 'new') is based on a single number, representing total activation (or 'familiarity', or 'degree of match'). In the present research we ask whether activation is increased by extra items added to memory (i.e. by list-length) and whether activation is increased by stronger or repeated items added to memory (i.e. by list-strength). The answers provide critical constraints for modelers of memory.

To help in understanding the experiments and the models, we review briefly the way in which performance in recognition tasks is measured. The subject studies a list of items and is then given items to judge as 'old' or 'new'. Responses of 'old' to items from the list (targets) are correct and termed 'hits'; responses of 'old' to new items (distractors) are incorrect and termed 'false alarms'. These data are usually related to theory in the following way: The subject is assumed to base the recognition judgment upon a measure representing 'familiarity', 'summed activation', 'degree of match', or some similar

statistic. It is assumed that targets have a distribution of familiarity with a somewhat higher mean than the distribution for distractors, as shown in Figure 1. A given trial results in a sample from either the target or distractor distribution: An old response is given if the sampled value is higher than a subject-chosen criterion. Thus, the hit probability is the area under the target curve above the criterion, and the false alarm probability is the area under the distractor curve above that criterion.

If an experimental manipulation is carried out on a list (such as increasing its length or strength), it is of course possible that the subject will adjust the criterion to a new position, changing the hit rates and false alarm rates. For this reason, in most studies the concern is not with the absolute levels of hits and false alarms but rather a comparison between them that can be used to measure sensitivity of performance. The measure **d'** is theoretically independent of the placement of the criterion and is defined as the difference between the means of the target and distractor distributions, divided by the (common) standard deviation:

$$d' = \frac{\mu_t - \mu_d}{\sigma}$$

This measure is simply calculated from the standard normal (z) transforms of the hit and false alarm rates.

In the present research, however, we are interested in the placements of the distributions and their movements, information not available from **d'.** We therefore developed an experimental procedure in which it is reasonable to expect the criterion to remain fixed over the conditions of interest. Under these circumstances, various models can be discriminated by the values of the hit and false alarm rates.

Experiments

We embedded categories of words in a single long list for study. Words from a given category (but never the prototype) were spaced randomly throughout the list, disguising the category structure. The length and strength of different categories were varied. Following presentation of the study list, items were randomly chosen and tested. The distractors that were tested included non-studied exemplars from each category, non-studied category prototypes and items from none of the categories. A studied item (target) from each category was also tested.

The experiment consisted of eight conditions: 4 length manipulations and 4 strength manipulations. The L-1, L-2, L-6, and L-10 conditions were comprised of 1, 2, 6, or 10 words chosen from a given category. Each of these words was presented once. The 'Pure' strength conditions consisted of 6 words, each of which was presented the same number of times; for different categories the number of repetitions were 1, 2, and 3. Unlike these Pure conditions, the 'Mixed' strength condition consisted of 6 words of which 2 were presented once, 2 were presented twice and 2 were presented three times. The selected words from all 8 conditions were randomly placed into a study list of 335 presentations, and presented for 3 seconds each. The test list consisted of 170 words for which the subject gave 'old' or 'new' judgments.

There were two separate types of word categories used in the experiment: Semantic and orthographic-phonemic. For example, the semantic category of prototype 'butterfly' contained the words 'moth', 'nectar', 'fragile', 'cocoon', 'monarch', 'flutter', 'metamorphosis', 'dragonfly', 'flitting', 'caterpillar', and 'camouflage'. In general, the semantic categories consisted of long words with relatively low natural language frequency (both prototypes and exemplars); the exemplars were all related to the prototype but not so closely that the presentation of an exemplar would likely call to mind the prototype. An example of an orthographic-phonemic category is that formed for the prototype 'sip'; it contained the words 'tip', 'lip', 'hip', 'sir', 'sin', 'sit', 'dip', 'rip', 'six', 'big', and 'fib'. In general, an orthographic-phonemic category contained short words with relatively high natural language frequency; the exemplars all shared vowels with the prototype, and one but not both of the starting and ending consonant clusters.

Predictions

Some patterns predicted for our study by typical models are illustrated in Figure 1. It is

important to note that the predictions are somewhat different from those for studies in which variables are manipulated one list at a time. In list studies, manipulations are often predicted to alter the variance of the distributions of activations. However, for essentially all models in which activations of all items participate in the resultant distributions, the mixing of so many categories of different types means that the variances of the distributions are not predicted to differ noticeably for different conditions.
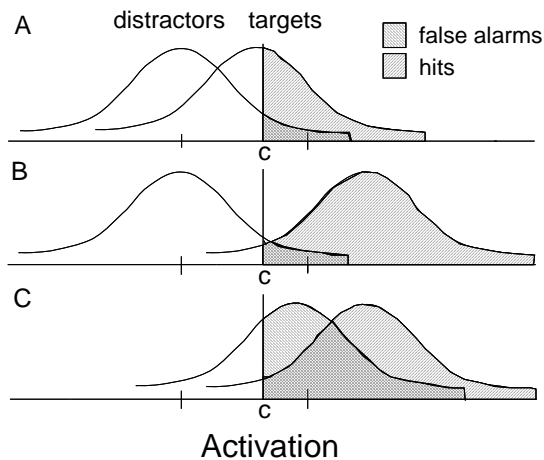

Figure 1. Probability distributions for activation due to targets and distractors. The panels represent predictions for experimental manipulations for various models (see text).

Given this, panel B illustrates the situation when a target is tested of increased strength relative to panel A, but where all items other than the target have unchanged strength-- the target distribution increases; false alarms are unchanged, hits rise, and **d'** rises (virtually all models). The category-length predictions vary with the model. For models in which non-target items contribute activations with greater than zero mean (e.g. the SAM model of Gillund and Shiffrin, 1984; the Matrix model of Pike, 1984), panel C illustrates the situation-- both distributions increase; hits and false alarms increase, and **d'** does not change. However, for models in which non-target items contribute zero mean activations (e.g. the TODAM model of Murdock, 1982; the CHARM model of Metcalfe, 1985; various feed-forward connectionist networks), the distributions do not move. This situation is given by panel A-- hits, false alarms and **d'** do not change.

For virtually all models, the category-strength predictions are at least qualitatively the same as the category-length predictions just discussed. That is, repetitions of an item should affect other items more or less as would an equivalent number of new items. However, Shiffrin, Ratcliff, and Clark (1990) discussed two alternative models. Their <u>differentiation</u> model utilized a tradeoff (discussed below) that causes summed activation to remain constant; this situation is represented by panel A. The category-length predictions are still those illustrated in panel C. The other model discussed by Shiffrin et al (1990) posited that both distributions increase with category strength, so panel C would illustrate these models' predictions for both length and strength. (For a list experiment this last model would predict a variance increase for length but not strength, but in the present category study, the variance differences "wash out").

Finally, we can consider the case in which <u>all</u> list items are increased in strength (in effect confounding the effect of target strength and the effect of strength of other category items). All models predict the same patterns as they do for strength of *other* items in the category (which for most models are also the predictions for category length), with the exception that the target distribution should increase, increasing the hit rates.

Results and Discussion

The results for performance (**d'**) were as follows: Increasing strength led to a sizable increase in performance, increasing length led to a slight decrease in performance, and there was no appreciable difference in performance between the mixed and pure conditions for the 1, 2, and 3, times presented items. The latter result indicates that category strength had little effect, since performance for an item of fixed strength (repetitions) is being compared when the other category members are varied in strength (repetitions). This finding is consistent with many earlier list studies (e.g. Ratcliff, Clark, & Shiffrin, 1990; Murnane & Shiffrin, 1991). As contrasted with the earlier list studies, however, all these **d'** data are consistent with the predictions of almost all models.

These performance results were expected, but the goal of the present research involved the separate hit and false alarm results. We assume that our procedure led the subjects to use a single criterion, regardless of the category, or category type, being tested. The pattern of results below is certainly consistent with this position.

Figure 2 shows the effect of increasing the strengths of all items in a category (solid lines). As predicted by all models, hit rates for targets rise; $t(888) = 4.871$, $p < .001$. Most importantly, the distractor false alarms clearly do not rise with strength of category, whereas the prototype false alarms show an upward trend that does not reach statistical significance.
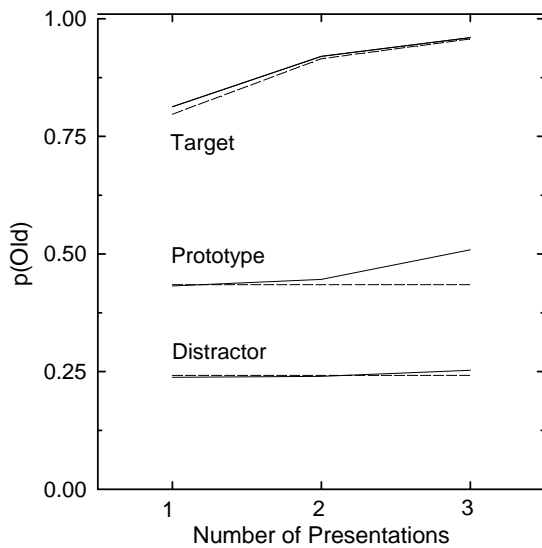

Figure 2. Probability of responding old for targets, prototypes, and distractors, as a function of the number of repetitions within a category. The solid lines are observed data and the dotted lines are predicted data.

The fact that the distractor false alarm rates did not rise with strength of category is consistent with the differentiation version of SAM, as well as those models that predict no shifts in the distributions for both category strength and category length (e.g. TODAM, CHARM, and some feed-forward connectionist nets).

The overall increased false alarm rate for prototypes would be predicted by almost all models on the basis of similarity; the prototypes should, on the average, be more similar to the words presented within the category than is the average distractor from that category. If the slight increase in the prototypes with strength is real, it would require explanation. It may be that subjects occasionally think of the prototype during the study list (especially when the words are repeated many times), and this occasionally leads the prototype to behave as a target at test.

In summary, the key result is the flat distractor function as strength increases.

Figure 3 shows the effect of extra items in the category. The probability of responding 'old' rises with category length for targets, $t(1184) = 2.54$, $p < .05$, for prototypes, $t(1184) = 6.840$, $p < .001$, and for distractors, $t(1184) = 3.357$, $p < .001$. Of the models consistent with the strength results, only the differentiation version of SAM predicts this pattern (since the other models predict similar patterns for length and strength). Thus the critical finding here is the contrast with the strength results. This contrast is reinforced by the category-strength effect that we turn to next.
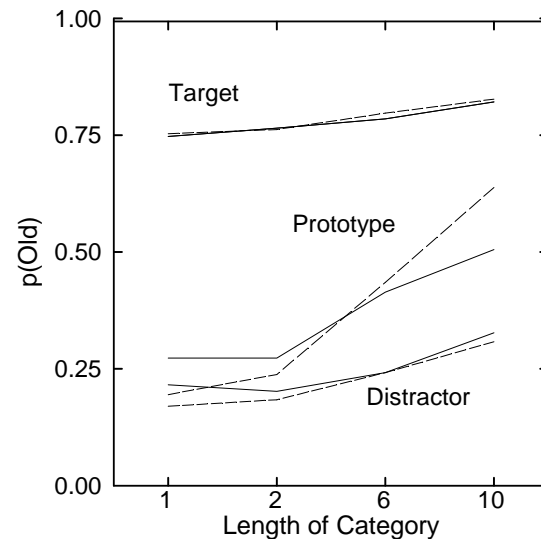

Figure 3. Probability of responding old for targets, prototypes, and distractors for categories of differing length. The dotted line represent predictions and the solid lines observed results.

Figure 4 depicts the effect of the strength of other items in a category by comparing the mixed and pure conditions for tested items of a given strength. None of these mixed/pure differences were significant, and all were small. Once again, the key point is the contrast with the length results. Although some models

predict no effect of category strength, they also predict no effect of category length. Of the models we have considered and know of, only the differentiation version of the SAM model predicts this pattern.
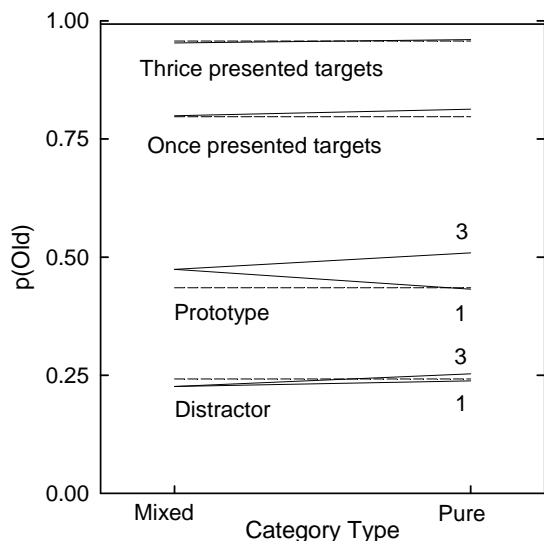

Figure 4. Probability of responding old for three times presented targets, once presented targets, prototypes, and distractors across the Mixed and Pure conditions. For both the prototypes and distractors there are two Pure points, representing tests from categories of differing strengths.

Modeling

To model the data, a simplified version of SAM was employed (see also Murnane and Shiffrin, 1991). In the SAM model introduced by Shiffrin et al (1990), it is assumed that all repetitions of an item are stored in the same trace; whenever an item is repeated, the new trace is appended to the pre-existing trace. At recognition, familiarity is determined by summing activations over all the traces. Activation of a given trace by two cues, a test item and a context cue, is posited to be the product of the separate activation tendencies for these two cues. An increase in repetitions for an item surely increases the activation tendency for the context cue. However, according to the differentiation hypothesis, an increase in the strength of a trace will cause it to mismatch more strongly the (different) test item. Thus, the activation tendency of the item cue will decrease. The model assumes that there is an

approximate tradeoff of these two opposing factors. Hence activation of trace i will not change with the repetitions of item i, when item j and the context cue are used as memory probes. Of course, when item i is activating trace i (tests of a target), then differentiation does not operate, and activation rises with strength. Finally, it is assumed that the standard deviation of activation of an image is proportional to the mean activation. The activations of all images are summed and compared to the criterion in order to make a decision.

The effect of similarity is dealt with in the following manner: Items outside the category of the probe have one level of activation, and items within the category have another. Items within the category when the probe is a proto-type have a third level of activation.

For computational ease, parameter estimation was performed through the fitting of predicted z-scores to the z-transforms of the hit and false alarm rates. The following is an example of the predicted formula for a target test from a Pure-2 category:

$$Z = \frac{C_r - [N_o S_o + N_i S_i + S_2]}{\sqrt{\alpha [N_o S_o^2 + N_i S_i^2 + S_2^2]}}$$

where C_r is the criterion, N_o is the number of out-of-category traces, S_o is the out-of-category strength, N_i is the number of different in-category traces, S_i is the in-category strength, S_2 is the strength of a twice repeated item that matches the probe, and α is a proportionality constant. The same type of formula applies to all cases; one must simply use, where needed, the remaining parameters: S_1 (once presented target strength), S_3 (three times presented target strength), and S_p (the in-category prototype strength). It is clear from the figures that the resultant predictions capture the main features of the data.

Conclusion

Under the assumption, partially supported by the data, that the subjects in our study utilize the same recognition criterion for all test items, a number of results concerning activation can be

drawn: 1) Increasing the number of presentations of an item causes greater activation when that item is tested; 2) Increasing the number of presentations of some items in a category does not cause more activation when some other item in that category is tested (either target or distractor); and 3) Additional items presented in a category cause more activation when any item in that category is tested (either target or distractor). These results are consistent with just one of the models we know of-- the differentiation version of the SAM model proposed by Shiffrin et al (1990).

It is interesting to note that the differentiation version of SAM was proposed to explain the lack of an effect upon **d'** of strengthening other items in a list for recognition. This lack of list-strength effect for recognition result was used to argue against models that are sufficiently composite in nature to produce storage interference. The present results further constrain the class of models that might be proposed for recognition memory: When an item is tested, increasing the number of presentations of some other item does not increase activation (i.e. familiarity, degree of match), even though additional items do cause such an increase.

References

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1-67.

Metcalfe Eich, J. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review, 92,* 1-38.

Murdock, B. B., Jr. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609-626.

Murnane, K., & Shiffrin, R. M. (1991). Word repetitions in sentence recognition. *Memory & Cognition, 19(2),* 119-130.

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review, 91,* 281-294.

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16,* 163-178.

Shiffrin, R. M., Ratcliff, R., & Clark, S. (1990). The list-strength effect. II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16,* 179-195.