Journal of Experimental Psychology:
Learning, Memory, and Cognition
1993, Vol. 19, No. 4, 799-812

Copyright 1993 by the American Psychological Association, Inc.
0278-7393/93/$3.00

An Evaluation of the Total Similarity Principle: Effects of Similarity on Frequency Judgments

Caren M. Jones and Evan Heit

Memory models that embody the *total similarity principle* (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988; Murdock, 1982; Ratcliff, 1990) assume that frequency judgments reflect the total similarity of a test item to stimuli that have been studied. In 4 experiments, subjects estimated the frequencies of target words that had been presented in the context of varying numbers of semantically similar words. In a fifth experiment, subjects made forced-choice relative frequency judgments. The results of these experiments supported 1 prediction of total similarity models: Presenting similar words will increase rather than decrease frequency judgments of target words. However, a second prediction of these models was not supported. In particular, similar-word presentations had no effect on the judged frequencies of target words that had not been shown.

Despite the apparent differences between various theories of memory, many of these theories share deep similarities. In particular, an important class of models, which we call *total similarity models*, assumes that frequency judgments and recognition judgments for a stimulus are monotonically related to the total similarity of the test item to the stimuli in memory. Although these models differ mathematically, the point is general. The more similar items that have been stored in memory, the higher the total similarity will be and the higher the frequency estimate will be for the test item. Likewise, as the total similarity increases, subjects will be more likely to recognize the test item as "old." Our approach in this article is to derive and empirically test implications of this total similarity principle.

The Total Similarity Models

We first explain how different memory models embody the total similarity principle.[1] Exemplar models such as Hintzman's (1988) MINERVA 2, Medin and Schaffer's (1978) context model, and Nosofsky's (1988) generalized context model have been applied successfully to frequency judgment tasks (e.g., Hintzman, 1988; Nosofsky, 1988) and recognition tasks (e.g., Estes, 1986; Heit, in press; Hintzman, 1988). These multiple-trace models assume that each stimulus results in a separate memory trace. When a test stimulus is presented for recognition or frequency estimation, it is compared with each trace, resulting in many individual similarity measures. These similarity measures are then summed into a total similarity score, which is assumed to be monotonically related to both frequency estimates and the probability of recognition. Thus, exemplar models make frequency and recognition judgments according to the total similarity principle.

Like the exemplar models, Gillund and Shiffrin's (1984) search of associative memory model (SAM) assumes that distinct events lead to separate memory traces. Recognition judgments are based on a familiarity measure, which is the sum of a test stimulus's associative strength to each memory trace. Because the associative strength between similar memory traces is greater than the associative strength between dissimilar memory traces, SAM is a total similarity model for recognition judgments. Although the SAM model has not been applied to frequency judgments, Gillund and Shiffrin (1984) have suggested that the familiarity measure could be used for that purpose. In the interest of developing and testing general principles, we assume that total similarity may be used for both recognition and frequency judgments, even if all of the models have not been applied to frequency judgments.

Composite vector models such as Murdock's (1982) theory of distributed associative memory (TODAM) and Metcalfe's (1985) composite holographic associative recall model (CHARM) assume that memory traces are overlaid into a single composite vector. However, the encoding and

Caren M. Jones, Department of Psychology, Stanford University; Evan Heit, Department of Psychology, University of Michigan.

Preparation of this article was supported by a National Science Foundation (NSF) Graduate Fellowship to Caren M. Jones, a National Institute of Mental Health Individual Postdoctoral Fellowship to Evan Heit, and an NSF Grant 91-10245 to Douglas Medin. We wish to thank Judy Florian, Robert Greene, Douglas Hintzman, Frances Kuo, Richard Marsh, Douglas Medin, Joshua Rubinstein, Holly Taylor, David Thau, and an anonymous reviewer for comments on earlier versions of this article; Kim Anderson, Mark Keavney, Paul Price, and Seema Shastri for assistance in running subjects; Chris Kortge for discussion of connectionist models; and J. E. Keith Smith for statistical advice.

Correspondence concerning this article should be addressed to Caren M. Jones, Department of Psychology, Stanford University, Jordan Hall, Building 420, Stanford, California 94305, or to Evan Heit, who is now at the Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, Illinois 60208.

[1] Related common characteristics of these models have also been noted by Hintzman, Curran, and Oppy (1992) and Shiffrin, Ratcliff, and Clark (1990).

retrieval operations preserve information about the identities and frequencies of presented stimuli. These models make recognition judgments by using the test stimulus as a retrieval cue, then evaluating the match value between the cue and what is retrieved. This match value increases with the number of times that the test stimulus, or similar stimuli, had been stored in the composite vector. The probability of recognition is assumed to be monotonically related to this match value; thus, the composite vector models are total similarity models. Presumably, these models would also make frequency judgments using the match value.

Two recent connectionist network models (Kortge, 1990; Ratcliff, 1990) also implement the total similarity principle. These models assume that stimulus presentations lead to changes in a network of associations. A test stimulus leads to a retrieved output vector, and the degree of match between the test stimulus and the output vector is used to make a recognition judgment. As in the composite vector models, more presentations of the tested stimulus, or similar stimuli, will lead to a greater match and a greater probability of recognition. Kortge (1990) has also shown that this match value can successfully predict the results of frequency judgment experiments. However, we do not claim that all connectionist models, or even all variations of the Kortge and Ratcliff models, are total similarity models because connectionist models are flexible enough to implement other principles.

Predictions of Total Similarity Models for Frequency Judgments

Although the total similarity principle is fairly general and is implemented in various ways, it still leads to testable predictions that are central to all of the members of the class of total similarity models. We focus on two predictions that are derived from the total similarity principle. First, total similarity models predict that the frequency judgments for a stimulus will be increased, rather than decreased, by the presentations of similar stimuli. For two different stimuli, A1 and A2, which have been presented an equal number of times, the frequency judgment for A1 will be greater than the judgment for A2 if other stimuli, B, similar to A1, have also been presented. The models make this prediction because the total similarity of A1 to what is in memory will include similarities to A1 traces as well as similarities to B traces. Another possibility is that in some cases, the judgment for A2 will be greater than for A1, and a model described in the next section makes this other prediction.

Second, total similarity models predict that the frequency judgment for a stimulus, A, will be increased by presentations of a similar stimulus, B, regardless of the number of A presentations. Although this prediction may be correct in some situations, we focus on the special case in which A has not been presented. Even in this case, total similarity models predict that occurrences of the similar stimulus B will increase the judgment for A because both memory traces for A and memory traces for B contribute independently to the total similarity when A is tested. Another possible outcome, not predicted by total similarity models, is that in some cases B presentations will not affect the judgment of A. We show in the next section that another memory model can make this alternate prediction.

Other Memory Models and Other Predictions

We have not found many theories of memory that are not total similarity models. Underwood (1969) proposed that presentation frequencies are encoded directly as attributes of a memory trace. In a frequency-counter model such as this, a frequency or recognition judgment for a stimulus is made by retrieving the attribute that denotes the number of presentations of that stimulus. Although Underwood (1969) did not predict how similarity would affect this process, it is plausible that the similarity of a test stimulus to other memory traces would sometimes lead a subject to retrieve the frequency counter for the wrong stimulus. Suppose that stimuli A1 and A2 have both been presented many times and that a stimulus, B, which is similar to A1, has been presented a few times. According to this conception of a frequency-counter model, sometimes when the frequency of A1 is estimated, the frequency counter for B will be mistakenly retrieved. On the average, the frequency estimate for A1 will be lower than the frequency judgment for A2 because sometimes testing with A1 will lead to retrieval of the counter for a low-frequency stimulus, B. Thus, this frequency-counter model is not a total similarity model; it predicts the opposite of the first prediction of total similarity models.[2]

Another model that is not a standard total similarity model is actually a variant of the SAM model. In the *differentiation* version of SAM (Shiffrin, Ratcliff, & Clark, 1990), judgments are still made using a familiarity measure determined by the summed similarity of a test stimulus to traces in memory. However, this model also allows similarity relations between stimuli to change with experience. For example, even if two stimuli, A and B, are initially highly similar, a great amount of study may make it easier to discriminate between them, in effect decreasing the similarity between A and B. Thus, presentations of B will not necessarily increase the frequency judgment for A because increased experience with B may actually decrease the total similarity of A to what is in memory. The differentiation version of SAM is highly flexible because the predictions of this model depend on the rate of differentiation. This model could be fitted to results that are consistent with the two predictions of total similarity models described earlier, but it could also be fitted to many results that violate the predictions.

[2] It is also plausible that these errors may be made during encoding, when the frequency counters are incremented. However, this possibility would still lead to the same predictions. Suppose that sometimes when A1 is presented, the counter for the similar stimulus, B, is incremented instead of the counter for A1. As long as the presentation frequency of A1 is greater than the frequency of B, then the frequency counter for A1 will be less than the counter for A2, for which there are no errors at encoding.

Previous Research Bearing on the Predictions of Total Similarity Models

Several studies (Begg, Maxwell, Mitterer, & Harris, 1986; Jacoby, 1972; Leicht, 1968; Shaughnessy & Underwood, 1973) have shown that situational frequency judgments of words are increased by the presentation of similar or semantically related words. However, these experiments do not provide a strong test of the first prediction of total similarity models because they used restricted ranges of presentation frequencies (0–5 at most). Thus, these studies do not rule out the frequency-counter theory prediction that higher presentation frequencies will be decreased in the presence of similar, low-presentation frequency words. Likewise, most of these studies do not permit assessment of the second prediction of total similarity models because they did not test subjects on nonpresented words that were similar to presented words. However, Leicht (1968) reported a small influence of presenting semantic associates on frequency estimates for nonpresented words but did not state whether this effect was statistically reliable. This result is also hard to interpret because it was averaged over several conditions, including similar words as well as dissimilar words that were thematically related.

Recent experiments by Hintzman, Curran, and Oppy (1992) provided a more systematic examination of similarity effects on frequency judgments. They presented either vacation slides or nouns over a wide range of frequencies (0–25 in one experiment). Subjects then made frequency estimates for the original stimuli and for distractor items that were easily confused with the original stimuli but had not been presented. For the vacation slides, the distractors were mirror images of the presented photos; for the nouns, the distractors differed from the original stimuli in terms of whether they had the suffix *s*. Hintzman et al. found that frequency judgments for the distractors were increased by the presentation of similar stimuli, so on the surface these results are consistent with total similarity models. However, frequency estimates for the unpresented, easily confused distractors had a bimodal distribution: About half of the judgments were zero, and about half were clustered near the actual frequency of the presented stimulus to which it was similar. The bimodal distribution is not predicted by total similarity models.

The results of Hintzman et al. (1992) do not bear directly on the first prediction of total similarity models considered in this article. Testing this prediction depends on comparing the direction of similarity effects on judgments for stimuli presented a different number of times, but Hintzman et al. only evaluated similarity effects on stimuli presented 0 times. Our second prediction concerns the presentation of similar stimuli on judgments for stimuli that had not been presented. Hintzman et al. did find such an effect for easily confused stimuli, but their result does not rule out the possibility of a different result for stimuli that are similar but less extremely so. Total similarity models predict that the frequency judgment for a target stimulus will also be increased by repetitions of a moderately similar stimulus.

Overview of Experiments

For Experiments 1 and 2, total similarity models and the frequency-counter model make conflicting predictions. When a target stimulus has been presented many times, what is the effect of presenting a smaller number of similar stimuli on judgments of the target stimulus? The total similarity principle predicts that presenting similar items will always increase the judgments of the target stimulus, but the frequency-counter model predicts that in this case, the frequency judgment for the target stimuli will be diminished. Experiments 3, 4, and 5 tested the second prediction of total similarity models. Do presentations of similar stimuli consistently increase the frequency judgment for a target stimulus, as predicted by the models, or are there some cases in which presenting similar stimuli has no effect? In particular, these experiments examined the possibility that the effects will be different for presented and unpresented target stimuli.

Experiment 1

Method

Subjects. Subjects were 24 Stanford University undergraduates who participated in partial fulfillment of a course requirement. No subject participated in more than one experiment or pilot study.

Materials. Two sets of stimulus words were created. The similar words were eight kinds of fish. The control words were eight words selected from eight different categories (Battig & Montague, 1969) and intended to be dissimilar to one another and to the fish stimuli. The two sets of words were matched on the range (fish = 50–216; control = 47–245) and mean (fish = 144; control = 152) of their category production frequencies (Battig & Montague, 1969) and on the range (fish = 3–9; control = 3–8) and mean (fish = 5.8; control = 5.4) of their lengths, in letters. An additional 10 words were chosen according to the same criteria as the control words and were used as filler words. The filler words were used to minimize primacy and recency effects on the similar and control words.

Pilot data collected from 44 San Jose State University undergraduates confirmed the similarity manipulation. On a 10-point scale, with 1 indicating *completely dissimilar* and 10 indicating *identical*, the mean pairwise similarity rating among the fish was 5.8, whereas the mean rating among the control words was 2.1. The mean similarity rating between the two word types was 2.3.

For each subject, one word of each type was randomly assigned to each of the presentation frequencies 0–7. The similar and control words were presented in a different random order for each subject, with the constraint that repetitions of a word be separated by at least 3 other words. Altogether, 66 stimuli were presented: 5 filler words, then 2 words at each of the frequencies 0–7, then another 5 filler words.

Procedure. The subjects were instructed to pay attention to the word displays for an unspecified memory test to be given afterward. Sixty-six words were then displayed one at a time, centered on the monitor of an IBM PC-compatible computer, at a rate of 3 s per word. Immediately following the presentation of all of the stimuli, subjects made absolute frequency judgments for each of the 16 similar and control words. Subjects were reminded that some words might not have been presented at all, and any nonnegative integer was accepted as a response. The words were tested in a different order for each subject.

Results

The mean frequency estimates are shown in Figure 1 as a function of presentation frequency and word type. The positive slopes of the lines show that subjects were sensitive to the presentation frequencies of both types of stimulus words: Judged frequency increased with presentation frequency. More important, the occurrence of words similar to the judged word affected the frequency estimate. The mean estimates for the fish were higher than those for the control words at all presentation frequencies, on the average by .61 judgment units. An analysis of variance (ANOVA) on the estimates including the within-subjects variables of word type and presentation frequency confirmed these observations. There was a main effect of presentation frequency, $F(7, 161) = 226.2, p < .001, MS_e = 2.22$, a main effect of word type, $F(1, 23) = 36.3, p < .01, MS_e = 2.71$, and no reliable interaction between the two variables ($p > .1$). Table A1 in the Appendix shows the mean and standard deviation for each condition.[3]

Discussion

This pattern of estimates clearly favors the total similarity models over the frequency-counter model. There was no indication in the data that higher frequencies of fish were diminished relative to control words, as the frequency-counter model predicts. Rather, the estimates for fish words were consistently greater than estimates for control words across presentation frequencies, as predicted by models that embody the total similarity principle.

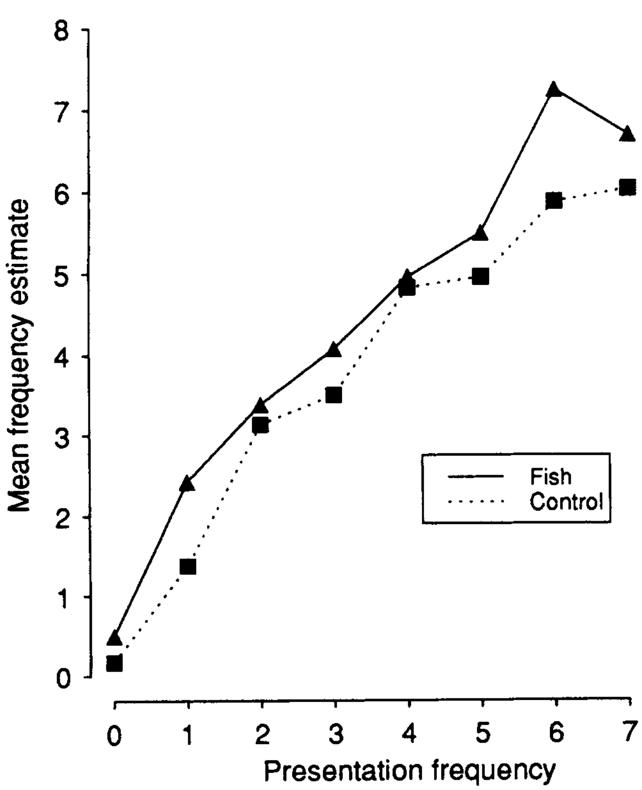

Figure 1. Estimated versus presentation frequency for the similar and control words of Experiment 1.

Figure 1 also suggests two effects that become important, and are tested more strictly, in Experiments 3, 4, and 5. First, the overestimation effect seems attenuated when the target word was presented 0 times. Despite being similar to 7 other words, presented a total of 28 times, the estimate for the similar word presented 0 times was only .33 units more than the estimate for the control word shown 0 times. Perhaps the similarity effect was mediated by the number of target presentations. Second, the similar word presented 7 times was judged to be slightly less frequent than the similar word presented 6 times. Perhaps the overestimation effect was also attenuated when a particular item was presented enough times that it became distinct from other items.

Experiment 2

Experiment 1 demonstrated that the occurrence of words similar to the target stimuli increased the frequency estimates to the targets. It may be argued, though, that there was something special about the category of similar stimuli presented. Despite the matching of words in the similar and control conditions, the results of Experiment 1 might have been affected by other differences between the similar and control words. Experiment 2 was intended to replicate Experiment 1 with a wider variety of stimulus categories and a larger range of presentation frequencies. Furthermore, in this experiment we drew stimuli for the similar and control conditions from the same pool of words to ensure that the two conditions would be comparable.

Method

Subjects. Subjects were 37 University of Michigan and 16 Stanford University undergraduates who participated in partial fulfillment of a course requirement.

Materials. Eight categories of stimulus words were used: fish, furniture, fruit, sports, musical instruments, vehicles, metals, and types of weather. From each category, 10 high-frequency exemplars were selected from Battig and Montague's (1969) norms. Four University of Michigan graduate students worked together to choose 7 exemplars from each category that minimized between-category similarity. For example, *lemon* was eliminated from the fruit category because it could also be considered a kind of vehicle. Thus, the result of this procedure was eight categories

[3] In the first four experiments reported here, the standard deviations of particular conditions were directly proportional to their means. This is a standard result for frequency judgments. Transforming the subjects' responses onto a logarithmic scale reduced the heterogeneity of variance but, with one exception, did not change the outcomes of the analyses reported here. A logarithmic transformation of the data in Experiment 3 did eliminate the otherwise significant interaction. It is common for transformations to eliminate interactions (Smith, 1976), but because of the ease of interpretation of our response scale, and the general lack of difference between the transformed and untransformed analyses, we report our results on the basis of untransformed data.

of 7 exemplars each, with words within a category being similar to each other but relatively dissimilar to words from the other categories. Ten filler words were also selected from other Battig and Montague categories.

Each subject saw all 7 exemplars from one particular category (the critical words) and 1 randomly selected word from each of the remaining categories (the control words), for a total of 14 stimulus words. The critical category rotated by subject number, and each category was presented to 6–8 subjects. One critical word and 1 control word were shown at each of the presentation frequencies 0, 2, 4, 6, 8, 10, and 12. Assignment of the frequencies to the words was randomized for each subject, as was the presentation order, except for the constraint that repetitions of a word were separated by at least 3 other words. There were also 5 filler words shown at the beginning of the list and 5 filler words at the end, for a total of 94 word presentations.

Procedure. The procedure was identical to that in Experiment 1, except that subjects saw 94 word presentations and were tested on 14 words.

Results

Figure 2 shows the mean frequency estimates for the critical and control stimuli as a function of presentation frequency. Table A2 shows the mean and standard deviation for each condition. Because the estimates did not differ across the eight critical categories, $F(7, 45) < 1$, the data are collapsed across the categories in Figure 2 and in the analyses that follow. As in Experiment 1, the positive slopes of the lines indicate that subjects were sensitive to presentation frequency. More important, the occurrence of similar stimuli affected the frequency judgments: At most presentation frequencies, the estimates for the critical words fell above those for the control words.

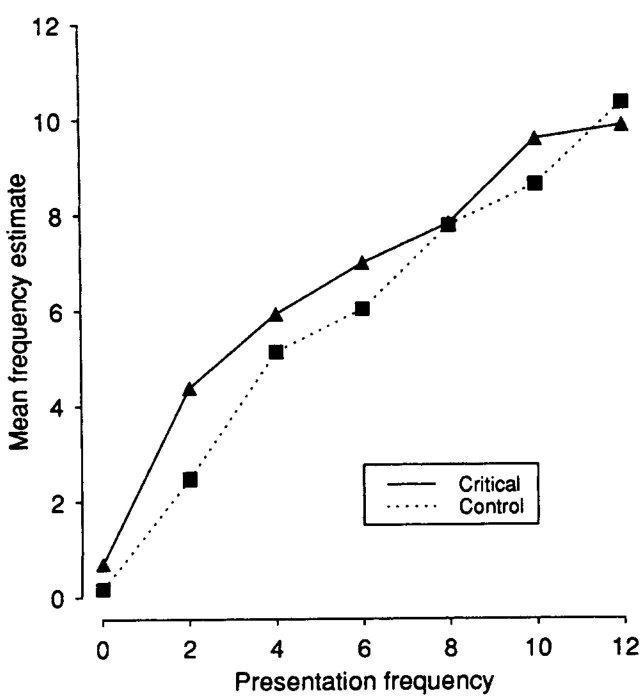

Figure 2. Estimated versus presentation frequency for the similar and control words of Experiment 2.

An ANOVA including the variables of presentation frequency and word type (critical vs. control) confirmed these observations. There was a main effect of presentation frequency, $F(6, 312) = 130.53$, $p < .001$, $MS_e = 9.08$, indicating that the frequency estimates increased with presentation frequency, and a main effect of word type, $F(1, 52) = 11.57$, $p < .01$, $MS_e = 7.22$, indicating that the category exemplars were given higher estimates than the control words. Unlike in Experiment 1, there was a weak Frequency × Word Type interaction, $F(6, 312) = 2.80$, $p < .05$, $MS_e = 5.40$. The surprising result that control judgments were greater than critical judgments at presentation frequency 12 was not reliable; Tukey's W test would require a pairwise difference of 1.72 between cells for a difference at the .05 level (Tukey, 1953, cited in Ott, 1988, pp. 446–449).

Discussion

These data replicate those of Experiment 1, again favoring the total similarity models over the frequency-counter model. There was no indication that estimates for higher presentation frequencies of the category words were lower than corresponding estimates for control words, as the frequency-counter model predicts; the apparent difference at presentation frequency 12 was not statistically reliable. It is not surprising that the results were somewhat noisier at higher presentation frequencies because the standard deviations of responses were higher. In addition, for critical words presented 12 times, the number of similar-word presentations was actually the least, compared with the number of similar-word presentations for the other critical words. Therefore, total similarity models would predict a somewhat smaller similarity effect at higher presentation frequencies. Even if the difference at presentation frequency 12 were reliable, the frequency-counter model could not explain these data. This model predicts that the critical and control frequency estimate functions should cross at the midpoint of the range of presentation frequencies (in this case, when the presentation frequency was 6). For frequencies smaller than that midpoint, the critical words should be judged as being more frequent than the control words, whereas for frequencies larger than that midpoint, the critical words should get lower estimates than the control words. Clearly, the data do not fit those predictions. Rather, the category words obtained higher judgments overall than did the control words, on the average by .67 judgment units. Note that this difference between conditions found in Experiment 2 is consistent with that found in Experiment 1 (.61 judgment units).

Experiment 3

In this experiment we examined the second prediction of total similarity models: that even when a test stimulus has not been presented, frequency estimates for it will increase if similar stimuli are presented. Pairs of semantically similar words, such as *creek–stream* and *robbery–burglary*, were used as stimuli. The use of similar pairs allowed us to extensively vary the frequencies of the target and similar words.

Method

Subjects. Subjects were 31 University of Michigan undergraduates who participated in partial fulfillment of a course requirement.

Materials. Sixteen pairs of similar words were obtained in the following manner. Twenty-four pairs of similar words were chosen from different categories of Battig and Montague (1969). Sixteen University of Michigan undergraduates rated the pairs on similarity, and we chose the 16 most similar pairs to serve as stimuli. In addition, 10 filler words were chosen from other Battig and Montague categories.

Each word was assigned a presentation frequency of 0, 3, 6, or 9. Frequency assignments were controlled so that there would be two word pairs with each of the paired target–similar-word frequencies 0–0 (neither target nor similar word shown), 0–3 (target not shown, similar word shown 3 times), 0–6, 0–9, 3–0, 3–3, 3–6, and 3–9.[4] The frequency assignment was rotated across subjects so that each word pair occurred with each frequency assignment nearly equally often (1 additional subject would have completely balanced the design). In addition, the assignment of frequencies to words within a pair was counterbalanced across subjects so that each word in a pair was the target word about equally often.

There were a total of 126 stimulus presentations: 12 words with a frequency of 3, 4 words with a frequency of 6, 4 with a frequency of 9, and 10 filler words displayed once each. The remaining 12 stimulus words had a frequency of 0 and were not shown. The order of the critical word presentations was randomized for each subject, with the constraint that repetitions of a word be separated by at least 3 other words.

Procedure. The procedure was identical to that in Experiment 1, except that subjects viewed 126 word presentations and were tested on 32 stimulus words.

Results

Figure 3 shows the frequency estimates for the target words as a function of target presentation frequency and similar-word presentation frequency. Once again, subjects showed sensitivity to differences in target presentation frequency. The crucial result is that similar-word presentations affected the estimates for target words presented 3 times but had no effect on target words presented 0 times.

An ANOVA including the variables of target frequency (0 or 3) and similar-word frequency (0, 3, 6, or 9) confirmed these observations. There were reliable main effects of both target-word frequency, $F(1, 30) = 115.7$, $p < .001$, $MS_e = 13.45$, and similar-word frequency, $F(3, 90) = 7.33$, $p < .001$, $MS_e = 1.09$. In addition, the interaction between variables was reliable, $F(3, 90) = 4.88$, $p < .01$, $MS_e = 1.02$, showing that similar-word frequency affected the Target Frequency 0 conditions differently from the Target Frequency 3 conditions.

Surprisingly, estimates for the presented target words increased as the similar words' frequency went up to 6 but then decreased as the similar words' repetitions increased to 9. Both the increase and the decrease were reliable (Tukey's $W = 0.70$, $p < .05$).

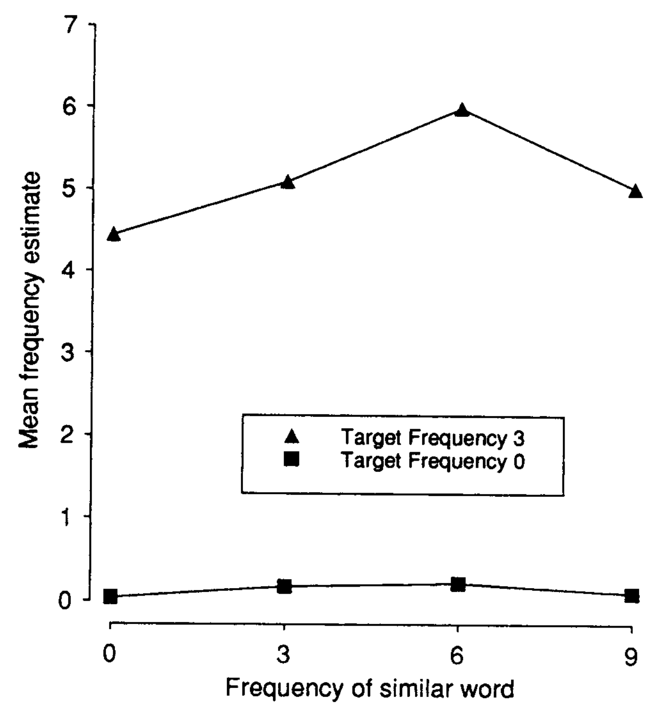

Figure 3. Estimated frequency of the target words as a function of presentation frequencies of the target and similar words of Experiment 3.

Discussion

The lack of effect of similar-word presentations on nonpresented test stimuli is a clear violation of the second prediction of the total similarity models. According to these models, similar-word presentations should increase the judgments for both presented target words and nonpresented target words. The failure of the second total similarity prediction was investigated further in Experiment 4.

The nonmonotonic effect of similar-word frequency on frequency judgments for presented words cannot be explained by the total similarity models either. The models predict a monotonic relation between the number of similar-word presentations and the frequency judgment for the target word. These models cannot explain why, for target words presented 3 times, frequency estimates were lower when 9 similar words were presented than when 6 similar words were presented. This result is suggestive of a differentiation process attributable to increased experience, such as the pro-

[4] Subjects estimated the frequency of each stimulus word, regardless of its conceptual status as a "target" or "similar" word. Therefore, subjects also estimated target words having frequencies of 6 and 9 with similar-word frequencies of 0 and 3. Because target words with frequencies of 6 or 9 had only limited numbers of similar words presented, we excluded these estimates from the data analysis, although they are included in Table A3. Another consequence of this design was that we collected twice as much data for the 0–0, 0–3, 3–0, and 3–3 conditions as for the 0–6, 0–9, 3–6, and 3–9 conditions.

posal by Shiffrin et al. (1990). This nonmonotonic effect was also investigated further in Experiment 4.

Experiment 4

Experiment 4 differed from Experiment 3 in that we used a larger range of similar-word presentations (up to 18) and included a recognition judgment condition. These changes were intended to bring the two surprising results from Experiment 3 into sharper focus. First, we hoped to replicate and extend the null effect of similar words on frequency estimates for nonpresented words by increasing the number of similar-word presentations. This null effect would be even more surprising when a larger number of similar stimuli are in memory. Furthermore, we tried to increase the generality of this result by also showing it in a recognition task. Second, we expected that the larger range of similar-word presentations would provide a good opportunity to replicate the nonmonotonic effect found in Experiment 3 because differentiation of memory traces should have a even larger effect with increasing frequencies.

Method

Subjects. Subjects were 81 San Jose State University undergraduates who participated in partial fulfillment of a course requirement. Each subject was assigned to either the frequency judgment condition or the recognition condition.

Materials. The stimuli were the same as those used in Experiment 3. Two word pairs were assigned each of the paired target–similar-word frequencies 0–0, 0–6, 0–12, 0–18, 3–0, 3–6, 3–12, and 3–18.[5] The assignment of word pairs to frequencies was rotated by subject number so that each word pair appeared about equally often with each paired frequency. Also, the assignment of frequencies to words within a pair was counterbalanced across subjects so that each word was the target word in a pair about equally often.

A total of 178 presentations were shown: 8 words with a frequency of 3; 4 words with frequencies of 6, 12, and 18; and 10 filler words shown once each. The remaining 12 stimulus words had a frequency of 0 and were not shown. Excluding the 10 filler words, the 168 stimulus presentations were randomly ordered for each subject, with the constraint that repetitions be separated by at least 1 other word.

Procedure. The procedure was the same as that used in Experiment 1, except for the following changes. Subjects observed 178 word presentations. Forty-two subjects then performed a recognition test on the 32 stimuli. For each word, they were instructed to press "1" to indicate that they had seen the word during the presentation and to press "0" to indicate that they had not seen the word. The remaining 39 subjects performed a frequency judgment task, just as in the previous experiments. Except for the response, the recognition test was identical to the absolute frequency judgment task.

Results

The mean frequency estimates for the target words are presented in Figure 4 as a function of target presentation frequency and presentation frequency of the similar words. Figure 5 shows the analogous data for the recognition judgments: The probability of judging a target word to be "old"

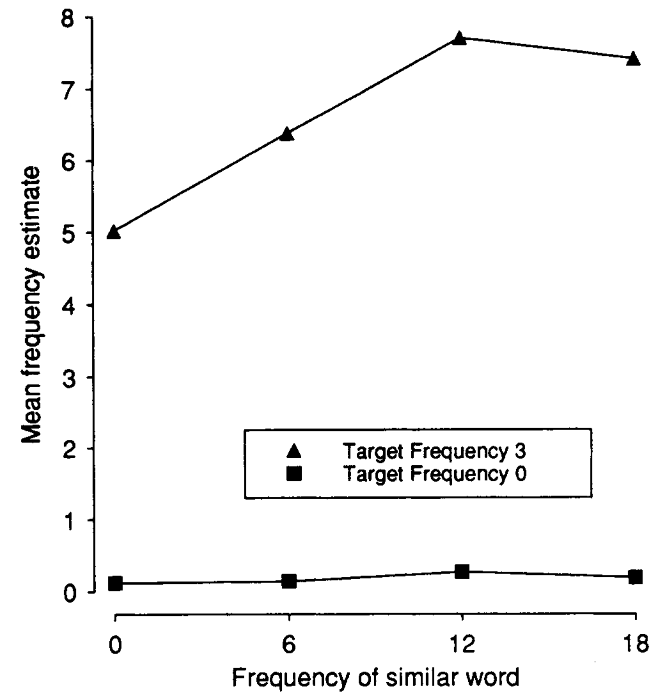

Figure 4. Estimated frequency of the target words as a function of presentation frequencies of the target and similar words of Experiment 4.

is plotted as a function of presentation frequency (0 or 3) and frequency of the similar word. Cell means and standard deviations are shown in Table A4.

Consider first the frequency data in Figure 4. There was a dramatic replication of the null similarity effect for nonpresented words, even with 18 similar-word presentations. No evidence was found for the prediction that similar words would influence frequency estimates for nonpresented words. Similarity did have a large positive influence on the estimates for presented stimuli, however, as indicated by the positive slope of the line for words shown 3 times. As in Experiment 3, there was a noticeable flattening of the estimates at the highest frequency of the similar words. Unlike Experiment 3, though, there was no reliable decrease in the judgments.

An ANOVA on the frequency data indicated highly reliable main effects of presentation frequency, $F(1, 38) = 106.2$, $p < .001$, $MS_e = 30.4$, and similar-word frequency, $F(3, 114) = 9.25$, $p < .001$, $MS_e = 3.33$, as well as their interaction, $F(3, 114) = 6.72$, $p < .001$, $MS_e = 3.78$.

Now consider the recognition data in Figure 5. Subjects were highly accurate: The overall false-alarm rate was only 4%. The occurrence of similar words did not affect recognition, as the lack of positive slope indicates. An ANOVA

[5] As in Experiment 3, we collected frequency estimates for all of the presented words. We did not include the target frequency 6, 12, or 18 in the analyses because the similar-word frequencies were not systematically varied. All of the estimates are described in Table A4.

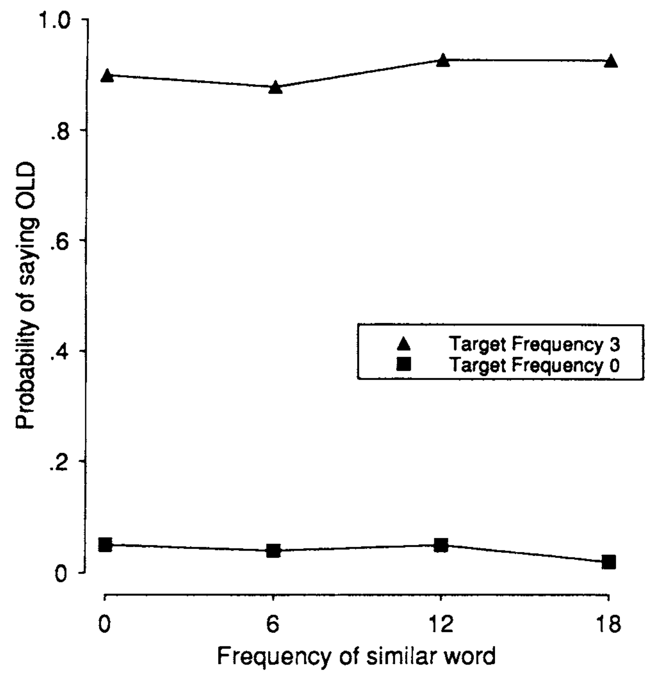

Figure 5. Probability of calling the target words "old" as a function of presentation frequencies of the target and similar words of Experiment 4.

conducted on only the recognition data confirmed these observations. There was a reliable main effect of presentation frequency, $F(1, 41) = 1279.9, p < .001, MS_e = 0.05$. Similar-word frequency and the interaction between the two variables were not found to be reliable (Fs < 1).

The differences, if any, between the recognition and frequency judgments for the nonpresented items are of interest: Were these judgments affected in the same way by the experimental manipulations? An ANOVA on both the recognition and frequency estimate data for just the nonpresented words indicated that the probability of giving a positive response (i.e., by either measure, saying that the word had occurred) was greater for the frequency estimates, $F(1, 79) = 6.34, p < .02, MS_e = 0.25$, but neither the effect of similar-word frequency nor the interaction was reliable (ps < 1). The larger probability of giving a positive response with frequency estimates than with recognition judgments was likely attributable to differences in how subjects used the two response scales. Range-frequency theory (Parducci, 1965) predicts that subjects will tend to rescale their judgments to cover the possible response categories. Because the subjects who made frequency judgments had many possible responses (i.e., the positive integers) for indicating that a stimulus had been presented, they were somewhat less likely than the recognition subjects to say that a stimulus had never been presented.

Discussion

The most important conclusion from Experiment 4, as well as Experiment 3, is that we have identified a case for which the second prediction of total similarity models fails. Presenting similar words influenced the frequency judgments for target words presented 3 times but had no effect on frequency judgments for nonpresented words. The critical result in the recognition condition was that, using the same stimuli, we again found no similarity effect on nonpresented words. This null effect serves as a further replication because total similarity models also predict that the false-alarm rate for nonpresented target words will increase when more similar words are presented. (Additionally, in an unpublished pilot study with a design similar to Experiment 4, similar words were presented up to 9 times. In this pilot study, too, similar-word presentations had no effect on frequency judgments or recognition judgments for words that had not been presented.) The low false-alarm rate obtained in Experiment 4 may suggest that our manipulations of similarity and presentation frequency were too weak. However, the obvious effects of similar words on presented targets, in all of the experiments reported so far, show that the similar stimuli were indeed similar enough to each other. In addition, 18 presentations of a similar word is a large number, and we doubt that presenting the similar word even more times would have changed the results of Experiment 4.

Note that unlike the data of Hintzman et al. (1992), the data in Experiments 3 and 4 were not bimodally distributed. Most subjects correctly identified nonpresented items as such, and the few who erred tended to give low frequency estimates rather than estimates clustering around the actual frequency of the similar presented item. It is possible that subjects in the Hintzman et al. studies were simply making discrimination errors. The stimuli in those studies were easily confused, much more so than our stimuli. Occasional failures to discriminate a nonpresented test stimulus from the similar presented stimulus led subjects to sometimes make frequency judgments for nonpresented stimuli that were close to the presentation frequencies for the similar stimuli. When their subjects successfully made this discrimination, however, they acted similar to our subjects: They showed no effect of presentations of similar stimuli on judgments.

It is also interesting to compare the judgments for nonpresented words between Experiments 1 and 2, and Experiments 3 and 4. The crucial difference between these pairs of studies is that in Experiments 1 and 2, several different similar words were presented, but in Experiments 3 and 4, a single similar word was presented repeatedly. Although the first pair of experiments seemed to show a similarity effect on nonpresented words (see Figures 1 and 2), the second pair of experiments (see Figures 3 and 4) showed no reliable effect, suggesting that the number of *distinct* similar words had an effect on nonpresented words, rather than simply the number of presentations of a single similar word.[6]

[6] Of course, it is difficult to draw a conclusion across different experiments. However, a recent recognition memory experiment by Huber, Ziemer, Shiffrin, and Marinelli (1992) led to the same conclusion. They found that increasing the number of distinct similar stimuli increased the false-alarm rate for a nonpresented target stimulus but that increasing the number of repetitions of the same similar stimulus had no effect on the false-alarm rate.

However, it is still not clear why repeated presentations of a single similar word would affect presented target words but not nonpresented target words.

How could total similarity models account for the null effect when target frequency is 0? One possible but post hoc account makes additional assumptions about the estimation scale. Total similarity models simply predict a monotonic relation between the total similarity level and the response scale, and the criteria or thresholds for which certain total similarity levels lead to particular frequency estimates are essentially free parameters. Suppose that these models additionally assume a high criterion for saying that an event occurred at least once. With this assumption, the total similarity would need to be fairly large before a frequency judgment of 1 would be made, so even 18 presentations of a similar word might still lead subjects to give a frequency judgment of 0 for a nonpresented target word.

This scaling explanation has substantial problems. Even with a strict criterion for saying that a word occurred once, we would still expect some small effect of similar words because different subjects would likely have different criteria. Also, the scaling function underlying this account is inconsistent with the rest of the data. There was a smaller increase in frequency judgments between 3 and 6 presentations of the target word than between 0 and 3 presentations, suggesting a logarithmic or negatively accelerated scaling function. However, this is inconsistent with a larger effect of similar-word presentations for targets presented 3 times than for targets presented 0 times. In Experiment 5 we obtained forced-choice judgments rather than numerical estimates so that interpretation of the results would not require assumptions about response scaling.

The other important conclusion for Experiment 4 is that the nonmonotonic effect of similar-word presentations on target words presented 3 times did not replicate.[7] We find the idea of differentiation attributable to experience appealing, but our findings for large numbers of similar words do not provide clear evidence for this effect.

Experiment 5

Experiments 3 and 4 demonstrated that absolute frequency judgments for nonpresented words were not affected by the presentation of similar words. Experiment 5 was intended to provide converging evidence for this lack of effect with a task that did not require assumptions about how subjects mapped their responses onto a numerical scale. As in Experiments 3 and 4, the stimuli were pairs of semantically similar words. In this experiment, however, subjects made forced-choice relative frequency judgments between pairs of words (e.g., *tuna–burglary*). For the critical test items, the words had been presented an equal number of times (0, 3, or 6) and differed only in the number of times their similar words had been presented (0, 3, 6, or 9). Thus, although the actual frequencies of the critical test words were the same, their total similarities differed.[8]

According to total similarity models, subjects are more likely to say that stimulus A1 is more frequent than stimulus A2 to the extent that the total similarity of A1 is greater (Gillund & Shiffrin, 1984; Hintzman, 1988; Kortge, 1990). For example, suppose that stimuli A1 and A2 have both been presented n times and that stimulus B, similar to A1, has been presented m times. The presentations of B lead the total similarity of A1 to be greater than that of A2, so subjects will tend to judge A1 as being more frequent than A2. These models also predict that the net effect of presenting B m times will be diminished as n increases (Hintzman, 1988, p. 531; Kortge, 1990, p. 768). These models make this prediction because the means and the variances of the total similarities of A1 and of A2 will increase as n increases, so the relative impact of the similarity of stimulus A1 to memory traces of B will decrease. In other words, if the total similarities of two target words differ by a constant amount, subjects' choices between these two words will be closer to chance performance when these two words have been presented many times. Thus, total similarity models predict that the similarity effect in this experiment will be *greatest* on nonpresented words.

Method

Subjects. Subjects were 74 University of Michigan undergraduates who participated for pay or course credit.

Materials. The stimuli were 24 pairs of semantically similar words, including the 16 pairs from Experiments 3 and 4. The remaining 8 pairs were chosen by asking an additional 26 University of Michigan undergraduates to judge the similarities between 30 pairs of words, each pair from a different Battig and Montague (1969) category. The 8 most similar word pairs from this set were used as stimuli, with the constraint that these additional words had to come from different Battig and Montague categories than the original 16 pairs of words. Two word pairs were assigned each of the paired target–similar-word frequencies 0–0, 0–3, 0–6, 0–9, 3–0, 3–3, 3–6, 3–9, 6–0, 6–3, 6–6, and 6–9. The assignment of word pairs to frequencies was rotated by subject number so that each word pair appeared about equally often with each paired frequency. Also, the assignment of frequencies to words within a pair was counterbalanced across subjects so that each word was the target word in a pair equally often.

A total of 190 presentations were shown: 14 words each with frequencies of 3 and 6, 6 words with a frequency of 9, and 10 filler words shown once each to absorb primacy and recency effects. The remaining 14 stimulus words had a frequency of 0 and were not shown. Excluding the filler words, the 180 stimulus presentations

[7] Another experiment, with a design similar to that of Experiment 3, also did not show the nonmonotonic result (Jones & Heit, 1991, Experiment 2). That study found that judgments for a target word presented 3 times strictly increased as the number of similar-word presentations increased from 0 to 9.

[8] Our rationale followed that of a classic study (Hintzman, 1969). That study investigated the effects of an experimental manipulation (massed vs. spaced presentations) by focusing on the case in which the manipulation may have the largest effect. For more than half of the forced-choice test questions, subjects were given two stimuli that had been presented the same number of times. Hintzman (1969) found a robust effect of massed versus spaced presentation on these test questions. Likewise, we intended to maximize the potential effect of similar words by focusing on test pairs for which the presentation frequencies were equal.

were randomly ordered for each subject, with the constraint that repetitions had to be separated by at least 3 other words.

There were 72 test pairs in four blocks of 18 items each. The blocks were conceptual only and were not apparent to the subject. The blocking was intended to ensure that particular words would be distributed evenly throughout the test phase. Within each block, there were 9 critical items for which the words had been presented an equal number of times, and 9 filler items in which the words differed in presentation frequency. The critical items were target–similar-word frequencies *X*–0 versus *X*–*Y*, where $X = 0$, 3, and 6 and $Y = 3$, 6, and 9. Recall that there were two words at each target–similar-word frequency. This meant that we could ask four critical test questions at each combination of frequencies without repeating any pairs of words. For example, if *tuna* and *burglary* were each assigned a target–similar-word frequency of 0–0, and *rain* and *baseball* were each assigned 0–3, then the four 0–0 versus 0–3 test items would be *tuna–rain, tuna–baseball, burglary–rain,* and *burglary–baseball.* One of the four items was assigned to each test block. The same structure also applied to the other kinds of critical items; thus, no critical test pair was ever repeated for a subject.

The filler items were chosen with two constraints: They had to differ in presentation frequency, and they had to balance out the testing of particular words such that each individual word was tested 3 times within each of two blocks (i.e., 6 times overall). The left–right ordering of the words was random for each test pair and for each subject, and the ordering of the items within each block was random for each subject. Finally, only one word from each pair of similar words appeared in the test phase (e.g., if *tuna* was tested, then *salmon* was not).

Procedure. The presentation phase was identical to the previous experiments, except that subjects saw 190 word presentations. For the testing phase, half of the subjects were asked to judge which word in each test pair had appeared more times during the presentation, and half of the subjects were asked to judge which had appeared fewer times, to ensure that the judgments overall were not biased by the direction of the question (see Hintzman & Gold, 1983). All subjects pressed "Q" to indicate the word on the left and "P" to indicate the word on the right.

Results

The data were collapsed across instruction condition (choosing the more or less frequent word). For clarity, we describe the data as if all subjects had been asked to choose the more frequent word. Figure 6 shows the data for the critical test items. The proportion of subjects choosing the word with the higher total similarity is shown for each level of actual frequency of the words and for each pairing of similar-word frequencies (i.e., for each pair of total similarity levels). Chance performance (i.e., a 50% proportion) would correspond to no effect of similar words, whereas a choice percentage greater than 50% would indicate a similarity effect. Figure 6 suggests that a similarity effect was obtained for words presented 3 or 6 times but that the choice proportions were close to chance for nonpresented words. Also, the similarity effect seemed to be slightly larger when similar words were presented 9 times compared with when similar words were presented 3 or 6 times.

An ANOVA was run on the critical items including variables of target-word frequency, similar-word frequency, and test block. There was a main effect of target-word frequency, $F(2, 146) = 3.50$, $p < .05$, $MS_e = .465$, indicating that the size of the similarity effect differed at different levels of target frequency. A planned contrast revealed that the choice proportions in the target frequency 0 conditions were reliably lower than the choice proportions in the target frequency 3 and 6 conditions, $F(1, 73) = 5.50$, $p < .05$, $MS_e = .44$). The main effect of similar-word frequency was not reliable, $F(2, 146) = 1.29$, $p > .2$, $MS_e = .26$, nor was the interaction between target-word and similar-word frequency ($F < 1$). There was a reliable main effect of test block, $F(3, 219) = 7.63$, $p < .001$, $MS_e = .338$. However, the differences between test blocks were not systematic or interpretable to us. The mean percentages of choices of the word with higher total similarity for the four test blocks were 58%, 49%, 64%, and 60%, respectively. The predictions of total similarity models about test-lag effects were not clear, and we had no other prior hypotheses, so we did not consider the test block variable further.

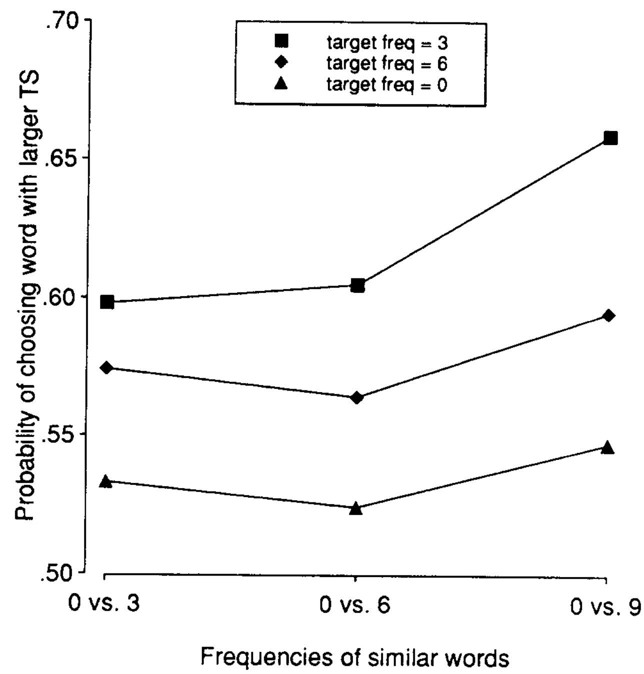

Figure 6. Proportion of judgments that the word with the higher total similarity (TS) was the more frequent word for each level of target frequency (freq) in Experiment 5.

To determine whether the choice proportions differed from chance, we examined the proportion of times that subjects chose the word with higher total similarity over the word with lower total similarity, for different levels of target word frequency. (The data were collapsed over the other variables, which did not interact with target frequency.) For the target words presented 3 or 6 times, subjects did reliably choose the word with the larger total similarity ($M = .62$), $t(73) = 4.47$, $p < .001$, $MS_e = .03$, and ($M = .58$), $t(73) = 3.05$, $p < .01$, $MS_e = .03$, respectively. When the target words had not been presented, however, subjects appeared to have chosen randomly. The probability of choosing the word with the larger total similarity was not greater than chance ($M = .53$), $t(73) = 1.60$, $p > .1$, $MS_e = .02$.

Finally, subjects answered a mean of 78% of the filler test items correctly, clearly above chance performance, $t(73) = 10.35$, $p < .001$, $MS_e = .03$. For completeness, we present the means for the three pairings of different target frequencies, although it must be cautioned that similar-word frequency was not counterbalanced for these filler test pairs. For target–similar-word frequencies 0–x versus 6–y, 0–x versus 3–y, and 3–x versus 6–y, the choice proportions were .84, .82, and .68, respectively. Because the filler questions were intended to balance the number of word presentations during the test phase rather than to test other hypotheses, the results for filler questions were not analyzed further. Nonetheless, the ordering of choice proportions for the three overall conditions is consistent with prior research on relative frequency judgments (e.g., Hintzman & Gold, 1983). We interpreted the relatively small difference between the 0 versus 6 proportion and the 0 versus 3 proportion simply as a ceiling effect.

Discussion

The most notable result of Experiment 5 is that it showed reliable effects of similar words on target words presented 3 or 6 times but no reliable effect of similar words on nonpresented target words. This result serves to replicate Experiments 3 and 4, in which we also found no effect of similar words on frequency judgments for nonpresented words. Moreover, this result is an extension of the previous findings in at least two ways: First, Experiment 5 showed that the similarity effect was reliably lower for nonpresented target words than for presented target words, in a case in which total similarity models predict a *greater* effect on nonpresented words. This finding, with relative frequency judgments, provides converging evidence that the second prediction of total similarity models is wrong. This finding adds to evidence provided by the null effects found for absolute frequency judgments in Experiments 3 and 4 and a pilot experiment, as well as recognition judgments in Experiment 4 and the pilot experiment. In these six different experiments or conditions of experiments, we found no statistically reliable similarity effects on nonpresented words, but we did find clear similarity effects on presented words. Of course, these results do not rule out the possibility that there might have been some tiny, statistically undetectable similarity effects on nonpresented words, but we find this proposal implausible, especially because Experiment 5 failed to find an effect where the total similarity models predicted that the effect would be largest.[9]

The second extension to our previous findings provided by Experiment 5 is that, as expected, we found a reliable similarity effect for target words presented 6 times. In Experiments 3 and 4 we did not systematically vary similar word presentations for target frequencies above 3, so the nature of the similarity effect on words presented more than 3 times was not clear. (Tables A3 and A4 show that it is difficult to draw conclusions about these conditions because the standard deviations of responses were fairly high.) Thus, Experiment 5 extends the generality of our findings of similarity effects. Figure 6 also suggests that the similarity effect was smaller for target frequency 6 than for target frequency 3, although this difference was not statistically reliable. Note that this difference would be predicted by the total similarity models, which assume that the effect of similar words will decrease as target frequencies increase.

Figure 6 suggests that the similarity effect increases with increasing numbers of similar words, but this trend was not statistically reliable. Again, the crucial finding from this experiment was that we did find similarity effects for presented targets but not for nonpresented targets. Total similarity models would also predict that the magnitude of the similarity effect would vary with the number of similar-word presentations, but the results of Experiment 5 do not allow us to draw conclusions about this second-order prediction.

General Discussion

Evaluating the Total Similarity Models

As a class, the total similarity models received mixed support from our results. Experiments 1 and 2 demonstrated that the occurrence of several different words that were semantically similar to a target word increased frequency estimates to that target. The overestimation was found to be roughly constant across the presentation frequencies 0–12, indicating that total similarity models account for the data better than does a frequency-counter model. Thus, the first prediction of total similarity models was supported: Frequency estimates were increased rather than decreased when similar stimuli had been observed.

However, we found a clear case in which the total similarity models' second prediction fails. Although Experiments 1 and 2 seemed to show that repetitions of several distinct similar words led to an increase in the judgment for a nonpresented target word, the later experiments consistently showed no effect of repeating a single similar word many times, contrary to the second prediction. Experiment 3 showed that subjects accurately estimated the frequency of target words that had not been presented, with no influence of similar-word presentations. Experiment 4 replicated the null effect of repeated similar stimuli, for up to 18 presentations, on nonpresented target words and indicated that the

[9] The lack of similarity effect on nonpresented words in Experiment 5 could be accounted for with the post hoc assumption that subjects perform forced-choice relative frequency judgments between two stimuli by first making numerical estimates of the absolute frequencies of each stimulus. The additional assumptions about numerical scaling from the Discussion section of Experiment 4 could then be applied. However, we know of no evidence supporting the assumption that people perform a relative frequency judgment by first performing two absolute frequency judgments. Additionally, the total similarity models that have been applied to relative frequency judgments (Hintzman, 1988; Kortge, 1990; see also Gillund & Shiffrin, 1984) do not make this assumption. Instead, these models predict that a relative frequency judgment is made directly from the total similarity scale, without an intermediate stage of numerical estimations of absolute frequency. Therefore, the existing total similarity models do predict a relatively large similarity effect on nonpresented words.

outcomes for recognition and frequency judgment tasks were the same. Experiment 5 ruled out response-scaling explanations for the null effect by replicating the result using a forced-choice method. These null effects could not have been caused by a simple lack of similarity between stimuli because presenting similar stimuli led to robust increases in judgments for target stimuli presented 3 times in Experiments 3 and 4 and to similarity effects for stimuli shown 3 and 6 times in Experiment 5.

Extensions to Total Similarity Models

Because total similarity models have been so successful on the whole, our negative findings with respect to the second prediction led us to propose that total similarity models should be improved rather than abandoned. We now describe a few possible extensions or modifications that could be implemented in total similarity models and evaluate whether they can account for our findings.

How could total similarity models account for the null effects on nonpresented words? One extension to total similarity models would be to assume that a second process occurs in which memory is searched for at least one memory trace that is a close match to the test stimulus. If there is no close match, then a "quick rejection" is made and the frequency estimate and probability of falsely recognizing the item become zero. This is much like Raaijmakers and Shiffrin's (1981, pp. 131–132) suggestion of how recognition judgments may be made in SAM. They argued that if the total similarity measure is not sufficiently high or low to immediately give a "new" or "old" judgment, then the memory traces are searched for a close match to the test probe. A recent study by Hintzman and Curran (1993), using a response–signal technique, has provided some evidence for such a content-matching process and for what its time course might be. However, this second process would have to be fleshed out considerably before it could be added to total similarity models. For example, it is not clear how subjects would decide what counts as a close match, particularly when some forgetting has occurred. Although total similarity models may be able to fit our results with such an extension, it is noteworthy that this extension requires the models to use other information instead of total similarity.

A second explanation for greater similarity effects at higher target presentation frequencies is that memory traces interact. Suppose that a subject is judging the frequency of *tuna* when both *tuna* and *salmon* have been presented several times. The *salmon* presentations may facilitate memory for *tuna* presentations by improving their encoding or retrieval. Thus, memory traces for *tuna* will have a greater impact on the total similarity, and the estimate to *salmon* will be increased. However, if *tuna* has not been presented, then *salmon* cannot affect *tuna* encoding or retrieval, and thus *salmon* has no affect on the judgment of *tuna*. Ross, Perkins, and Tenpenny (1990) have found evidence for such interactions, and Hintzman (1988) has already described an additional trace interaction process for MINERVA 2. Simply adding such a mechanism to current total similarity models will not succeed, though, because total similarity models still predict an influence of *salmon* presentations independent of any facilitation of *tuna* memory traces, even if *tuna* has not been presented.

A final way to help total similarity models explain the failure of the second prediction is to assume that the similarities between words are not fixed but vary because of experimental conditions such as presentation frequency. One such proposal, by Shiffrin et al. (1990), is that memory traces will become differentiated, and less associated with test stimuli, with more presentations. Their proposal provides a simple explanation for the null effects of similar stimuli that we found for nonpresented target stimuli. Suppose that the frequency of a stimulus, A1, is estimated after a similar stimulus, B, has been presented several times. The presentations of B could diminish the associative strength between A1 and B so that judgments for A1 would not be increased by its total similarity to memory traces for B. However, this account should predict the same lack of effect of B presentations regardless of the number of A1 presentations, but we found a robust effect of similar-word presentations when the target word had also been presented. Therefore, differentiation, as described by Shiffrin et al., is not enough to explain our complete pattern of results.

Another possible effect of presentations on similarity is that simply being presented in the same experimental context increases the similarity between words. Conversely, presenting one word but not another leads the two words to be less similar. The reason that presenting similar words did not affect judgments of nonpresented target words is that the words differed in a salient way: One had already been presented in the experiment and the other had not. Thus, the total similarity of a nonpresented target word was not affected by presentations of semantically related words.

Extending total similarity models with a more flexible concept of similarity, which may be influenced by both differentiation and differences in past contexts, can account for the results of Experiments 3, 4, and 5. However, these extensions may make the predictions of total similarity models too flexible and unconstrained.

Conclusion

Other recent reports also suggest that total similarity models have trouble with the specific details of the effects of repeated stimuli on frequency or recognition judgments. Hintzman et al. (1992) claimed that their results cannot be explained by total similarity models. Shiffrin et al. (1990) also discussed the results of repeated presentations of some stimuli on recognition judgments for other stimuli. Even when the stimuli are not especially similar to each other, total similarity models still predict that recognition memory for target stimuli will be worse when other stimuli have been presented many times. These models make this prediction because presentations of other stimuli will increase the variance of the total similarity of target stimuli to traces in memory, making it more difficult to distinguish between presented stimuli and nonpresented stimuli. However, subjects instead showed no such effect of repetitions on recognition memory or they showed an effect in the opposite direction. Finally, Florian (1992; Medin & Florian, 1992) and Nosofsky (1991) also presented results that pose problems for total similarity models' assumption that each presentation of a stimulus simply leads to a new encoding of that stimulus. These studies showed complex patterns for the effects of presenting different numbers of similar stimuli on recognition and classification judgments. Indeed, Florian found that in some cases, the total similarity of a target stimulus appeared to be decreased by increased presentations of similar stimuli. All of these researchers have proposed that total similarity models must either relax the assumption that each stimulus presentation leads to a separate encoding into memory or relax the assumption that the similarity relations between stimuli are independent of presentation frequency.

How are repeated presentations of a stimulus encoded into memory? How is the similarity between memory traces affected by frequency of occurrence? It is clear that to explain our frequency judgment results, as well as the results of other recent studies, total similarity models of memory must better address these fundamental questions.

References

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs, 80*(3, Pt. 2).

Begg, I., Maxwell, D., Mitterer, J. O., & Harris, G. (1986). Estimates of frequency: Attribute or attribution? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 496–508.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Florian, J. (1992). Frequency effects on categorization and recognition. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 826–831). Hillsdale, NJ: Erlbaum.

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

Heit, E. (in press). Modeling the effects of expectations on recognition memory. *Psychological Science.*

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *Journal of Experimental Psychology, 80,* 139–145.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Hintzman, D. L., & Curran, T. (1993). *Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall* (Tech. Rep. No. 93–1). Eugene: University of Oregon, Institute of Cognitive and Decision Sciences.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 667–680.

Hintzman, D. L., & Gold, E. (1983). A congruity effect in the discrimination of presentation frequencies: Some data and a model. *Bulletin of the Psychonomic Society, 21,* 11–14.

Huber, D. E., Ziemer, H. E., Shiffrin, R. M., & Marinelli, K. (1992). Does memory activation grow with list strength and/or length? In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 147–152). Hillsdale, NJ: Erlbaum.

Jacoby, L. L. (1972). Context effects on frequency judgments of words and sentences. *Journal of Experimental Psychology, 94,* 255–260.

Jones, C. M., & Heit, E. (1991, November). *Constraints on similarity effects for situational frequency judgments of words.* Paper presented at the 32nd Annual Meeting of the Psychonomic Society, San Francisco.

Kortge, C. A. (1990). Episodic memory in connectionist networks. In *Proceedings of the 12th Annual Meeting of the Cognitive Science Society* (pp. 764–771). Hillsdale, NJ: Erlbaum.

Leicht, K. L. (1968). Recall and judged frequency of implicitly occurring words. *Journal of Verbal Learning and Verbal Behavior, 7,* 918–923.

Medin, D. L., & Florian, J. E. (1992). Abstractions and selective coding in exemplar models of categorization. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of William K. Estes* (pp. 207–234). Hillsdale, NJ: Erlbaum.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

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

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

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27.

Ott, L. (1988). *An introduction to statistical methods and data analysis* (3rd ed.). Boston, MA: PWS-KENT Publishing Company.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285–308.

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology, 22,* 460–492.

Shaughnessy, J. J., & Underwood, B. J. (1973). The retention of frequency information for categorized lists. *Journal of Verbal Learning and Verbal Behavior, 12,* 99–107.

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

Smith, J. E. K. (1976). Data transformations in analysis of variance. *Journal of Verbal Learning and Verbal Behavior, 15,* 339–346.

Underwood, B. J. (1969). Attributes of memory. *Psychological Review, 76,* 559–573.

(Appendix follows on next page)

Appendix

Means and Standard Deviations of Frequency Estimates in Experiments 1–4

Table A1
Means and Standard Deviations of Frequency Estimates in Experiment 1

Presentation frequency	Fish M	Fish SD	Control M	Control SD
0	0.50	1.02	0.17	0.38
1	2.42	1.28	1.38	1.31
2	3.38	1.28	3.13	1.51
3	4.08	1.59	3.50	1.35
4	4.96	1.49	4.83	1.71
5	5.50	2.23	4.96	2.10
6	7.25	3.35	5.88	2.07
7	6.70	2.10	6.04	2.46

Table A2
Means and Standard Deviations of Frequency Estimates in Experiment 2

Presentation frequency	Similar M	Similar SD	Control M	Control SD
0	0.70	1.66	0.16	0.47
2	4.38	2.40	2.47	1.37
4	5.92	3.37	5.13	2.76
6	6.98	4.08	6.02	2.93
8	7.79	4.00	7.75	4.26
10	9.54	4.48	8.60	3.72
12	9.83	4.68	10.30	4.90

Table A3
Means and Standard Deviations of Frequency Estimates in Experiment 3

	Similar-word frequency							
	0		3		6		9	
Target-word frequency	M	SD	M	SD	M	SD	M	SD
0	0.04	0.15	0.18	0.52	0.23	0.84	0.10	0.40
3	4.44	2.32	5.10	3.13	6.00	3.45	5.03	2.37
6	7.39[a]	3.99	7.52[a]	3.73				
9	8.94[a]	4.27	8.31[a]	3.33				

[a] Not included in the statistical analysis because the design of the experiment was not fully factorial.

Table A4
Means and Standard Deviations of Frequency Estimates in Experiment 4

	Similar-word frequency									
	0		3		6		12		18	
Target-word frequency	M	SD	M	SD	M	SD	M	SD	M	SD
0	0.12	0.24			0.14	0.57	0.27	0.86	0.19	0.82
3	5.02	3.09			6.37	3.34	7.68	5.48	7.38	5.41
6	8.03[a]	4.72	9.17[a]	4.55						
12	11.94[a]	4.92	11.26[a]	5.44						
18	14.97[a]	5.79	13.42[a]	5.91						

[a] Not included in the statistical analysis. Because of a bad computer disk, 6 subjects' data were lost for these points.

Received July 29, 1991
Revision received October 26, 1992
Accepted November 1, 1992 ■